ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΔΙΟΙΚΗΤΙΚΗΣ ΜΕΤΑΡΡΥΘΜΙΣΗΣ & ΗΛΕΚΤΡΟΝΙΚΗΣ ΔΙΑΚΥΒΕΡΝΗΣΗΣ

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ & ΑΥΤΟΔΙΟΙΚΗΣΗΣ

ΙΝΣΤΙΤΟΥΤΟ ΕΠΙΜΟΡΦΩΣΗΣ

ΕΣΠΑ 2007-2013 — Ε.Π. ΔΙΟΙΚΗΤΙΚΗ ΜΕΤΑΡΡΥΘΜΙΣΗ — Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΠΙΜΟΡΦΩΤΙΚΟ ΠΡΟΓΡΑΜΜΑ

«ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΜΕ ΣΤΑΤΙΣΤΙΚΑ ΠΑΚΕΤΑ»

ΕΚΠΑΙΔΕΥΤΙΚΟ ΥΛΙΚΟ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1. Εισαγωγή στη Στατιστική 6
1.1 Οι Βασικές Έννοιες 7
1.2 Στατιστικά στοιχεία: Συλλογή οργάνωση και παρουσίαση 9
1.2.1 Στάδιο πρώτο: Συλλογή στατιστικών στοιχείων 10
1.2.2 Στάδιο δεύτερο: η οργάνωση των στατιστικών στοιχείων 10
1.2.3 Στάδιο τρίτο: η παρουσίαση των στατιστικών στοιχείων 11
1.3 Παραδείγματα 11
Παράδειγμα 1 11
Παράδειγμα 2 13

ΚΕΦΑΛΑΙΟ 2. Το SPSS και το περιβάλλον εργασίας του 15
2.1 Το SPSS και το περιβάλλον εργασίας του 16
2.1.1 Στατιστικά πακέτα και πεδία εφαρμογής τους - Το πρόγραμμα SPSS και δυνατότητές του 16
2.1.2 Περιγραφική Στατιστική – Στατιστική παρουσίαση και εξέταση δεδομένων με τη χρήση του SPSS – Πίνακες Συχνοτήτων 18
2.1.3 Το αρχείο αποτελεσμάτων του SPSS 24
2.1.4 Το αρχείο εντολών του SPSS 27

ΚΕΦΑΛΑΙΟ 3. Περιγραφική στατιστική I : πίνακες και διαγράμματα 31
3.1 Η έννοια του στατιστικού πληθυσμού 32
3.2 Η έννοια της παρατήρησης 33
3.3 Η έννοια της κατανομής συχνότητας 33
3.3.1 Η περίπτωση της ασυνεχούς μεταβλητής 33
3.3.2 Η περίπτωση της συνεχούς μεταβλητής 34
3.4 Η γραφική παρουσίαση της κατανομής 35
3.4.1 Το Ιστόγραμμα Συχνοτήτων 35
3.4.2 Η Πολυγωνική Γραμμή ή Πολύγωνο Συχνοτήτων 35
3.5 Η αθροιστική κατανομή συχνοτήτων: η έννοια των κατανομών «μικρότερη από» και «μεγαλύτερη από» 37
3.6 Η σχετική κατανομή συχνότητας και η σχετική αθροιστική κατανομή συχνότητας 39
3.7 Παραδείγματα 40
Παράδειγμα 1 40
Παράδειγμα 2 41
Παράδειγμα 3 42

ΚΕΦΑΛΑΙΟ 4. Περιγραφική Στατιστική II : μέτρα συμπύκνωσης των πληροφοριών 43
4.1 Οι παράμετροι τάσης 44
4.1.1 Ο Αριθμητικός Μέσος 44
4.1.2 Γεωμετρικός Μέσος 46
4.2 Οι παράμετροι θέσης 47

4.2.1 Η Διάμεσος 47
4.2.2 Τα Τεταρτημόρια, Δεκατημόρια, Εκατοστημόρια 49
4.2.3 Η Επικρατούσα Τιμή 51
4.3 Οι παράμετροι διασποράς 52
4.3.1 Το Εύρος Διασποράς 52
4.3.2 Η Μέση Απόκλιση 53
4.3.3 Η Διακύμανση και η Τυπική Απόκλιση 53
4.3.4 Ο Συντελεστής Μεταβλητικότητας 55
4.3.5 Οι Τυποποιημένες (Ανηγμένες) Διαφορές 55
4.4 Οι ροπές κατανομής συχνοτήτων 57
4.5 Οι παράμετροι ασυμμετρίας 57
4.6 Οι παράμετροι κύρτωσης 60
4.7 Παραδείγματα 61

Κεφάλαιο 5. Στοιχεία Πιθανοθεωρίας I 67
5.1 Εισαγωγή 68
5.2 Πείραμα τύχης, ενδεχόμενα, δειγματικός χώρος 69
5.3 Προσδιορισμός δειγματικού χώρου 71
5.4 Έννοιες, ιδιότητες και προσδιορισμός πιθανοτήτων 74
5.5 Τυχαίες μεταβλητές και συναρτήσεις πιθανότητας. 83
5.6 Βασικές παράμετροι τυχαίων μεταβλητών. 87

Κεφάλαιο 6. Στοιχεία Πιθανοθεωρίας II 92
6.1 Εισαγωγή 93
6.2 Διωνυμική κατανομή 93
6.2.1 Η έννοια του διωνυμικού νόμου πιθανότητας 93
6.2.2 Υπολογισμός διωνυμικών πιθανοτήτων 95
6.2.3 Παράμετροι και μορφή της διωνυμικής κατανομής 98
6.3 Κανονική κατανομή 103
6.3.1 Η έννοια της κανονικής κατανομής 103
6.3.2 Ιδιότητες της κανονικής κατανομής 103
6.3.3 Πιθανότητες και κρίσιμες τιμές κανονικών τυχαίων μεταβλητών 105

Κεφάλαιο 7. Δειγματοληψία και Κατανομές Δειγματοληψίας 110
7.1 Εισαγωγή 111
7.2 Περί δειγματοληψίας 112
7.2.1 Η έννοια του τυχαίου δείγματος 112
7.2.2 Σχέδια δειγματοληψίας 115
7.2.3 Σφάλματα δειγματοληπτικών ερευνών 120
7.3 Κατανομές δειγματοληψίας μονομεταβλητών πληθυσμών 120
7.3.1 Βασικές έννοιες κατανομών δειγματοληψίας 120
7.3.2 Κατανομή δειγματοληψίας του μέσου 124
7.3.3 Αναγκαία παρένθεση. Δειγματοληψία από κανονικούς πληθυσμούς: Βασικές έννοιες των κατανομών χ^2, t-Student, και F. 127

7.3.4 Κατανομή δειγματοληψίας της τυπικής απόκλισης 132
7.3.5 Κατανομή δειγματοληψίας της αναλογίας 133

Κεφάλαιο 8. Εκτιμητική και Κλασικοί Παραμετρικοί Έλεγχοι Στατιστικών Υποθέσεων 135
8.1 Εισαγωγή 136
8.2 Εκτιμητές και μέθοδοι εκτίμησης 136
8.2.1 Εκτιμητές και ιδιότητές τους 136
8.2.2 Μέθοδοι εύρεσης εκτιμητών πληθυσμιακών παραμέτρων 140
8.3 Εκτίμηση σε διάστημα εμπιστοσύνης 141
8.3.1 Έννοια και μεθοδολογία κατασκευής διαστημάτων εμπιστοσύνης 141
8.3.2 Εφαρμογές: Διαστήματα εμπιστοσύνης για το μέσο, την τυπική απόκλιση και την αναλογία 146
8.4 Έλεγχος στατιστικών υποθέσεων 152
8.4.1 Έννοια και μεθοδολογία κλασικού παραμετρικού ελέγχου υποθέσεων 152
8.4.2 Εφαρμογές: Έλεγχοι υποθέσεων για το μέσο, την τυπική απόκλιση και την αναλογία 162

ΚΕΦΑΛΑΙΟ 9. Ανάλυση Διακύμανσης κατά Παράγοντα 168
9.1 Εισαγωγή 169
9.2 Ανάλυση διακύμανσης με ένα κριτήριο 170
9.2.1 Το υπόδειγμα 170
9.2.2 Έλεγχος υποθέσεων 173
9.2.3 Κριτήρια πολλαπλών συγκρίσεων (post hoc tests) 178
9.2.3.1 Κριτήριο LSD 178
9.2.3.2 Κριτήριο Bonferroni 180
9.2.3.3 Κριτήριο Tukey HSD 181
9.2.3.4 Κριτήριο Scheffe 183

ΚΕΦΑΛΑΙΟ 10. Ανάλυση Παλινδρόμησης και Συσχέτισης 184
10.1 Εισαγωγή 185
10.2 Απλή γραμμική παλινδρόμηση 185
10.2.1 Εκτίμηση των παραμέτρων α και β 187
10.2.2 Σφάλματα εκτίμησης ή κατάλοιπα 191
10.2.3 Τυπικό σφάλμα εκτίμησης 193
10.2.4 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων της παραμέτρου β 194
10.2.5 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων της παραμέτρου α 198
10.2.6 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων του μέσου της κατανομής Y (Y_{x_i}) 200
10.2.7 Εκτίμηση μιας τιμής της Y (Πρόβλεψη) 203
10.2.8 Ανάλυση διακύμανσης 206
10.2.9 Συντελεστής προσδιορισμού (R) 210
10.3 Συσχέτιση 211
10.3.1 Συντελεστής γραμμικής συσχέτισης 212

10.3.2 Συσχέτιση και παλινδρόμηση 215
10.3.3 Έλεγχος υποθέσεων συντελεστή γραμμικής συσχέτισης 216
10.4 Πολλαπλή παλινδρόμηση 219
10.4.1 Υπόδειγμα πολλαπλής γραμμικής παλινδρόμησης 219
10.4.2 Εκτίμηση της εξίσωσης της πολλαπλής γραμμικής παλινδρόμησης 221
10.4.3 Τυπικό σφάλμα εκτίμησης 222
10.4.4 Ανάλυση διακύμανσης 222
10.4.5 Συντελεστής πολλαπλού προσδιορισμού 224
10.4.6 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων των παραμέτρων 225

ΚΕΦΑΛΑΙΟ 11. Μη παραμετρικοί έλεγχοι 227
11.1 Εισαγωγή 228
11.2 Κριτήριο ελέγχου των Kolmogorov – Smirnov 229
11.3 Κριτήριο ελέγχου του Wilcoxon (T) των σημασμένων διαβαθμίσεων 231
11.4 Κριτήριο ελέγχου των Mann-Whitney (U) 235
11.5 Κριτήριο ελέγχου του Wilcoxon (W) 239
11.6 Κριτήριο ελέγχου των Kruskall – Wallis (H) 243

ΕΡΩΤΗΣΕΙΣ 248

ΑΠΑΝΤΗΣΕΙΣ 260

ΠΑΡΑΡΤΗΜΑ I 268

ΠΑΡΑΡΤΗΜΑ II: Συσταδοποίηση 368

ΠΑΡΑΡΤΗΜΑ III 381

ΒΙΒΛΙΟΓΡΑΦΙΑ 389

ΗΛΕΚΤΡΟΝΙΚΕΣ ΔΙΕΥΘΥΝΣΕΙΣ 391

1. Εκπαιδευτική Ενότητα

• Εισαγωγή στη Στατιστική

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να γνωρίζει τη χρησιμότητα και πεδία εφαρμογής της Στατιστικής.
- Να γνωρίζει τους κλάδους της Στατιστικής.
- Να κατανοεί τις βασικές Στατιστικές έννοιες: πληθυσμός, δείγμα, μεταβλητές και τιμές τους, στατιστικά δεδομένα.
- Να αναγνωρίζει και να υλοποιεί μεθόδους συλλογής δεδομένων και να επισημαίνει τα στατιστικά σφάλματα.
- Να σχεδιάζει μία στατιστική έρευνα πρωτογενών δεδομένων.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Στατιστική
- Περιγραφική Στατιστική
- Στατιστικός πληθυσμός
- Στατιστικό Δείγμα
- Ποσοτική-Ποιοτική Μεταβλητή
- Συλλογή-Οργάνωση-Παρουσίαση Στατιστικών Στοιχείων

1.1 Οι Βασικές Έννοιες

Ο όρος Στατιστική προέρχεται από τη λατινική λέξη status που σημαίνει κράτος ή κατάσταση, επειδή ο αρχικός ρόλος της συλλογής και επεξεργασίας στοιχείων αριθμητικών δεδομένων ήταν να παρέχουν πληροφορίες στο κράτος σχετικά με οικονομικά ζητήματα[1]. Κατά τη διάρκεια του αιγυπτιακού, ελληνικού και ρωμαϊκού πολιτισμού η συλλογή στοιχείων γινόταν κυρίως για φορολογικούς λόγους αλλά και για να υπάρχει πληροφόρηση στο κράτος σχετικά με το στρατιωτικό του δυναμικό.

Ένας πρώτος ορισμός της Στατιστικής θα μπορούσε να την περιγράψει ως την επιστήμη που ασχολείται με τη συλλογή, οργάνωση, παρουσίαση και ανάλυση αριθμητικών στοιχείων ή χαρακτηριστικών με τη βοήθεια επιστημονικών μεθόδων και τεχνικών. Λέγοντας αριθμητικά στοιχεία ή χαρακτηριστικά αναφερόμαστε σε κάθε οικονομικό, επιχειρηματικό, κοινωνικό, δημογραφικό κ.λπ. φαινόμενο. Η επεξεργασία αυτών με τη βοήθεια της Στατιστικής οδηγεί σε συμπεράσματα που μπορούν να χρησιμοποιηθούν ως εισροές στα μοντέλα που εφαρμόζονται στη διαδικασία λήψης ορθών αποφάσεων στη διοίκηση. Η διαδικασία βελτίωσης των οικονομικών μονάδων απαιτεί, με βάση τις σύγχρονες θεωρίες, εφαρμογή του τρίπτυχου της Φιλοσοφίας Μάνατζμεντ-Επιστήμης της Συμπεριφοράς-Επιστήμης της Στατιστικής.

Η Στατιστική ασχολείται α) με την παρουσίαση και περιγραφή των πληροφοριών, β) την εξαγωγή συμπερασμάτων σχετικά με τους πληθυσμούς βασιζόμενη μόνο σε πληροφορίες που εξάγονται από ένα δείγμα και γ) την πραγματοποίηση αξιόπιστων προβλέψεων για μεταβλητές που μας ενδιαφέρουν. Τα υπό το στοιχείο α) αποτελούν την Περιγραφική Στατιστική ενώ τα υπό τα στοιχεία β) και γ) αποτελούν την Επαγωγική Στατιστική.

Λέγοντας Περιγραφική Στατιστική εννοούμε τη διαδικασία συλλογής, οργάνωσης και παρουσίασης των δεδομένων μας. Τα δεδομένα μπορούν να έχουν τη μορφή αριθμητικών στοιχείων ή τη μορφή χαρακτηριστικών. Η παρουσίασή τους έχει τη μορφή πινάκων και γραφημάτων. Τέλος, η ανάλυσή τους έχει τη μορφή επεξεργασίας τους με τη χρήση συγκεκριμένων στατιστικών περιγραφικών μέτρων τα οποία θα μορφοποιήσουν τα μη τακτοποιημένα με κάποιο συγκεκριμένο τρόπο δεδομένα μας σε τρόπο που να μπορούν να γίνουν αντιληπτά καταρχήν οπτικά και ακολούθως αριθμητικά. Η όλη χρήση της Περιγραφικής Στατιστικής μας δίνει μια εικόνα για το συγκεκριμένο δείγμα που εξετάζουμε και όχι για ολόκληρο τον πληθυσμό (εκτός εάν ολόκληρος ο πληθυσμός είναι αντικείμενο επεξεργασίας), τα συμπεράσματα στα οποία καταλήγει η ανάλυσή τους αφορά αποκλειστικά και μόνο το δείγμα αυτών των δεδομένων.

Στην περίπτωση της Επαγωγικής Στατιστικής τα συμπεράσματα του δείγματος με τη χρήση των νόμων των πιθανοτήτων ανάγονται στον συνολικό πληθυσμό. Αναφερόμενοι στον όρο Πληθυσμός ή Στατιστικός πληθυσμός εννοούμε το σύνολο των στοιχείων που δυνητικά εξετάζουμε και τα οποία παρουσιάζουν τουλάχιστον ένα κοινό χαρακτηριστικό, π.χ. το φορολογητέο εισόδημα που δηλώνουν οι Έλληνες φορολογούμενοι. Το σύνολο τον ατόμων που υποβάλλουν φορολογική δήλωση είναι ο πληθυσμός μας. Προκειμένου να εξάγουμε συμπεράσματα για το συνολικό πληθυσμό χρησιμοποιούμε την επαγωγική στατιστική. Η διαδικασία που ακολουθούμε είναι να εξάγουμε ένα δείγμα από αυτόν με συγκεκριμένη επιστημονική μεθοδολογία (δειγματοληψία) έτσι ώστε το δείγμα αυτό να είναι στατιστικό δείγμα και να χαρακτηρίζεται ως αντιπροσωπευτικό. Να μπορεί δηλαδή με βάση στατιστικές τεχνικές να χαρακτηριστεί ως μια μικρογραφία του συνολικού πληθυσμού.

---

[1] Λεξικό Γ. Μπαμπινιώτη, Α' Έκδοση, 2004

Επ' αυτού του δείγματος εφαρμόζονται στη συνέχεια οι μέθοδοι και οι τεχνικές που θα αναφερθούν στα επόμενα κεφάλαια ώστε να μπορέσουμε να καταλήξουμε σε συγκεκριμένα συμπεράσματα αναφορικά με το ποια είναι τα τυχόν ιδιαίτερα χαρακτηριστικά γνωρίσματα ή το ποιες είναι οι ιδιότητες που το διέπουν.

Η εφαρμογή της σωστής διαδικασίας λήψης δείγματος, κάτι που συνδέεται και με τη σχετική εμπειρία του ερευνητή αποτελεί προϋπόθεση για την εξαγωγή ενός ποιοτικού δείγματος. Το παραπάνω βέβαια σε συνδυασμό με την εφαρμογή της κατάλληλης μεθόδου ανάλυσης θα οδηγήσει στην εξαγωγή όσο το δυνατόν αξιόπιστων αποτελεσμάτων.

Μπορούμε στο σημείο αυτό να παραθέσουμε μια σειρά ορισμών που είναι απαραίτητοι για την πορεία της παρουσίασης.

Στατιστικός πληθυσμός είναι το σύνολο των μονάδων που απαρτίζουν τον πληθυσμό (π.χ. το σύνολο των υπαλλήλων του δημοσίου τομέα, το σύνολο των υπόχρεων σε υποβολή φορολογικής δήλωσης μιας χώρας).

Στατιστικό δείγμα είναι ένα υποσύνολο του στατιστικού πληθυσμού που εξάγεται με συγκεκριμένη στατιστική τεχνική (δειγματοληψία) προκειμένου να μας βοηθήσει στην ανάλυση του στατιστικού πληθυσμού και στην περιγραφή αυτού. Το δείγμα θα πρέπει να πληρεί συγκεκριμένα χαρακτηριστικά και ειδικότερα αυτά της αντιπροσωπευτικότητας και της αμεροληψίας. Θα πρέπει δηλαδή να έχει εξαχθεί με τρόπο αντικειμενικό και να εγγυάται ότι αποτελεί μια μικρογραφία του συνολικού πληθυσμού με όλες τις ιδιότητες που αυτός έχει.

Το χαρακτηριστικό ως προς το οποίο μελετάμε τον πληθυσμό είναι γνωστό ως μεταβλητή. Ο τρόπος με τον οποίο εκφράζουμε την κατάσταση της κάθε μεταβλητής είναι οι διάφορες τιμές της. Η τιμή μιας μεταβλητής μπορεί να εκφραστεί αριθμητικά, με κάποιο σύμβολο ή ακόμα και με έκφραση. Για τον συμβολισμό μιας μεταβλητής χρησιμοποιούμε κεφαλαία γράμματα (π.χ. η μεταβλητή X) ενώ για τον συμβολισμό των τιμών της χρησιμοποιούμε το αντίστοιχο μικρό γράμμα της μεταβλητής $x_1, x_2, x_3, \ldots, x_v$.

Δεδομένου ότι μεταβλητή μπορεί να είναι οποιαδήποτε ιδιότητα του πληθυσμού που μελετάμε άλλες μπορεί να είναι δυνατόν να εκφραστούν με αριθμούς και άλλες με σύμβολα ή εκφράσεις. Έτσι μπορούμε να διακρίνουμε περαιτέρω τις μεταβλητές ανάλογα με τον τρόπο έκφρασής τους σε ποιοτικές και σε ποσοτικές.

Η περίπτωση μιας ποιοτικής μεταβλητής περιγράφεται ως αυτή που δεν μπορεί να εκφραστεί ή να μετρηθεί αριθμητικά. Η αποτύπωσή της θα γίνει με ένα σύμβολο ή μια έκφραση. Ένα παράδειγμα ποιοτικής μεταβλητής είναι το επίπεδο μόρφωσης των υπαλλήλων του δημοσίου τομέα ή το χρώμα των ματιών των φοιτητών ενός πανεπιστημίου.

Η ποσοτική μεταβλητή από την άλλη μεριά είναι αυτή που επιτρέπει τη μέτρησή της με αριθμούς και έτσι η τιμή της είναι συγκεκριμένος αριθμός και όχι έκφραση όπως στην περίπτωση της ποιοτικής μεταβλητής. Ένα παράδειγμα μπορεί να είναι το ύψος των μηνιαίων απολαβών όλων των στελεχών ενός Υπουργείου. Μια περαιτέρω διάκριση των ποσοτικών μεταβλητών είναι αυτή που τις ορίζει ως συνεχείς ή ασυνεχείς. Έτσι ως συνεχείς μεταβλητές ονομάζουμε εκείνες που μπορούν να πάρουν όλες τις τιμές εντός ενός διαστήματος αναφοράς. Έτσι έχουμε μεταβλητές που προσδιορίζονται ως ποσοτικές συνεχείς μεταβλητές όπως για παράδειγμα το δηλωθέν φορολογητέο εισόδημα των πολιτών μιας χώρας. Η μεταβλητή αυτή μπορεί να έχει οποιαδήποτε τιμή από 0 έως

το άπειρο. Αντίστοιχα έχουμε μεταβλητές που προσδιορίζονται ως ποσοτικές ασυνεχείς. Πρόκειται δηλαδή για μεταβλητές που μπορούν να λάβουν ένα περιορισμένο αριθμό τιμών ενός διαστήματος αναφοράς. Ένα κλασσικό παράδειγμα είναι η ένδειξη από τη ρίψη ενός ζαριού. Στην περίπτωση αυτή οι τιμές που μπορεί να πάρει αυτή η μεταβλητή περιορίζονται στις {1, 2, 3, 4, 5, 6}. Πρόκειται δηλαδή για ένα πεπερασμένο αριθμό διακριτών τιμών. Άλλο χαρακτηριστικό παράδειγμα είναι ο αριθμός των φορολογούμενων πολιτών των κρατών μελών της Ε.Ε. Είναι προφανές ότι οι τιμές αυτής της ποσοτικής μεταβλητής δεν μπορεί παρά να είναι ακέραιες μονάδες αφού το σύνολο των πιθανών τιμών της {1, 2, 3, 4, 5, 6} είναι πεπερασμένο.

1.2 Στατιστικά στοιχεία: Συλλογή οργάνωση και παρουσίαση

Αναφερθήκαμε προηγούμενος στην έννοια της στατιστικής. Η στατιστική ως διαδικασία αφορά την επεξεργασία (στατιστική ανάλυση) δεδομένων προκειμένου να εξαχθούν συγκεκριμένα συμπεράσματα. Η βάση αυτή της στατιστικής ανάλυσης ή αλλιώς η πρώτη ύλη της διαδικασίας είναι τα στατιστικά στοιχεία. Αυτά αποτελούν τη βάση της στατιστικής ανάλυσης. Είναι φανερό ότι εάν έχουμε εξασφαλίσει υψηλή ποιότητα αυτών των στατιστικών στοιχείων μπορούμε να αναμένουμε αξιόπιστα συμπεράσματα. Το πρώτο βήμα λοιπόν στη στατιστική ανάλυση έχει να κάνει με τον τρόπο που θα πραγματοποιηθεί η συλλογή, οργάνωση και παρουσίαση των στατιστικών στοιχείων.

Καταρχήν θα πρέπει να ορίσουμε την έννοια των στατιστικών στοιχείων (στατιστικά δεδομένα). Τα δεδομένα μας προκειμένου να χαρακτηριστούν ως στατιστικά στοιχεία και όχι απλά θα πρέπει να είναι γεγονότα, συμβάντα ή ενδείξεις, τα οποία παρουσιάζουν ένα είδος μετρήσιμης σχέσης. Εάν αυτό το χαρακτηριστικό απουσιάσει και έχουμε απλά αυθαίρετα στοιχεία αυτά σε καμία περίπτωση δεν μπορούν να χαρακτηριστούν ως στατιστικά. Εάν για παράδειγμα έχουμε μια αυθαίρετη αρίθμηση μονάδων ενός πληθυσμού χωρίς να λαμβάνεται υπ' όψιν κάποιο κοινό χαρακτηριστικό ή ιδιότητα του πληθυσμού αυτά δεν αποτελούν στατιστικά στοιχεία. Στατιστικά στοιχεία όμως αποτελούν αυτά που προέρχονται από μια αρίθμηση μονάδων ενός πληθυσμού που λαμβάνει υπ' όψιν κάποιο κοινό χαρακτηριστικό ή ιδιότητα.

Η στατιστική ανάλυση στην οποία αναφερόμαστε είναι μια διαδικασία που περιλαμβάνει πέντε συγκεκριμένα στάδια:

Στάδιο πρώτο: Η συλλογή των στατιστικών στοιχείων.

Στάδιο δεύτερο: Η οργάνωση των στατιστικών στοιχείων.

Στάδιο τρίτο: Η παρουσίαση των στατιστικών στοιχείων.

Στάδιο τέταρτο: Η ανάλυση των στατιστικών στοιχείων και η εξαγωγή στατιστικών συμπερασμάτων.

Στάδιο πέμπτο: Η εφαρμογή των συμπερασμάτων του τέταρτου σταδίου στο σύνολο των μονάδων του πληθυσμού.

Παρακάτω παρουσιάζονται τα πρώτα τρία ενώ στο επόμενο κεφάλαιο τα άλλα δύο.

1.2.1 Στάδιο πρώτο: Συλλογή στατιστικών στοιχείων

Το πρώτο βήμα μας θα πρέπει να είναι ο προσδιορισμός ενός συγκεκριμένου προβλήματος που επιθυμούμε να μελετήσουμε. Το επόμενο βήμα είναι η συλλογή των στατιστικών στοιχείων που θα μας επιτρέψουν να έχουμε το απαραίτητο υλικό προκειμένου να προχωρήσουμε στην εφαρμογή της στατιστικής μας ανάλυσης του προβλήματος που μελετούμε. Η σχέση ποιότητας στατιστικών στοιχείων και συμπερασμάτων που εξάγουμε είναι μια αναλογική σχέση. Όσο πιο αξιόπιστα είναι τα στοιχεία μας τόσο πιο ορθά στατιστικά θα είναι τα συμπεράσματα που θα εξαγάγουμε. Μια συνήθης πρακτική αναφορικά με την αξιοπιστία των στοιχείων είναι η χρήση πρωτογενών δημοσιευμένων στοιχείων από αξιόπιστες στατιστικές βάσεις (π.χ. ΕΣΥΕ, ICAP, ΙΟΒΕ). Η απευθείας συλλογή στοιχείων είναι η λύση που θα πρέπει να χρησιμοποιείται στην περίπτωση που δεν υπάρχουν δημοσιευμένα στοιχεία για το αντικείμενο που ενδιαφερόμαστε να μελετήσουμε. Θα πρέπει όμως να εξασφαλίσουμε ότι τα στοιχεία μας θα έχουν όλα εκείνα τα χαρακτηριστικά που αναφέρονται παρακάτω στο σχετικό κεφάλαιο περί δειγματοληψίας. Θα πρέπει να πραγματοποιήσουμε συλλογή στοιχείων με βάση συγκεκριμένη διαδικασία και όχι στην τύχη. Σημειώνουμε εδώ ότι η έννοια της τυχαίας δειγματοληψίας είναι τελείως διαφορετική και απόλυτα επιστημονική και δεν πρέπει να συγχέεται με τη δειγματοληψία στην τύχη.

1.2.2 Στάδιο δεύτερο: η οργάνωση των στατιστικών στοιχείων

Εφόσον έχουμε στην κατοχή μας τα στατιστικά στοιχεία θα πρέπει να φροντίσουμε να τα παραθέσουμε με οργανωμένη μορφή σε σχέση με την άτακτη μορφή στην οποία βρίσκονται κατά το πρώτο στάδιο. Θα πρέπει δηλαδή τα στοιχεία μας να ταξινομηθούν και στη συνέχεια να παρατεθούν με τη μορφή πινάκων και διαγραμμάτων ώστε να έχουμε και την οπτική τους μορφή. Η διαδικασία αυτή πραγματοποιείται σε τρία τμήματα.

1.2.2.1 Τμήμα 1 : Διόρθωση των Στοιχείων

Στο τμήμα αυτό γίνεται ένα πρώτο ξεκαθάρισμα των στατιστικών μας στοιχείων στη βάση του εάν έχουν σχέση με την πραγματικότητα (εάν είναι δηλ. αληθοφανή), εάν έχουν λάθη (εάν συμφωνούν με τους γενικούς κανόνες και τις παραδοχές της θεωρίας) και εάν περιέχουν μεταξύ τους στοιχεία ψευδή (π.χ. σε μια δειγματοληψία η ποιότητα των στοιχείων εξαρτάται και από μη ελεγχόμενους παράγοντες όπως στην περίπτωση που ένας ερωτώμενος απαντάει ψέματα).

1.2.2.2 Τμήμα 2: Ταξινόμηση των Στοιχείων

Στο τμήμα αυτό η ταξινόμηση των στοιχείων γίνεται κατά πρώτο στη βάση του εάν έχουμε ποσοτικά ή ποιοτικά στοιχεία και κατά δεύτερο λαμβάνοντας υπόψη τον τόπο και το χρόνο.

1.2.2.3 Τμήμα 3: Πινακοποίηση των Στοιχείων

Στο τμήμα αυτό αφού τα στοιχεία μας έχουν ταξινομηθεί θα πρέπει να εισαχθούν σε πίνακες ώστε να αποκτήσουν μορφή που να τα κάνει πλήρως κατανοητά. Ένας πίνακας θα μπορούσε να κατέγραφε τα στοιχεία με βάση το έτος, ένας άλλος με βάση τον τόπο

λήψης (π.χ. νομός, συνοικία) και ένας άλλος με βάση τις ομάδες του πληθυσμού (π.χ. ηλικιακές ομάδες).

Οι επιλογές παράθεσης των στοιχείων σε πίνακες είναι πάρα πολλές. Μπορούμε να κατασκευάσουμε πίνακες απλής ή πολλαπλής ταξινόμησης, πίνακες αναφοράς ή ποσοστών, πίνακες διπλής εισόδου κ.λπ. Η επιλογή γίνεται με βάση το σκοπό της παρουσίασης.

Ένας τυπικός στατιστικός πίνακας περιέχει καταρχήν τον τίτλο του (π.χ. Πίνακας 1.3.2), κατά δεύτερον τον κύριο τίτλο των στοιχείων που παρουσιάζει (π.χ. Ακαθάριστο Εθνικό Προϊόν και Δημόσιο Χρέος των Χωρών Μελών του ΟΑΣΑ), τους τίτλους των στηλών του πίνακα (ΑΕΠ, ΔΧ), την πηγή (Πηγή: ΟΟΣΑ, Αύγουστος 2009), και τις τυχόν σημειώσεις του πίνακα (Σημειώσεις: Προσωρινά στοιχεία, Εκτιμήσεις κ.λπ.).

1.2.3 Στάδιο τρίτο: η παρουσίαση των στατιστικών στοιχείων

Κατά το στάδιο αυτό θα πρέπει τα οργανωμένα πλέον στοιχεία μας να λάβουν μια πιο αναλυτική και εύκολα αντιληπτή μορφή. Η μορφή αυτή είναι η διαγραμματική απεικόνιση των στοιχείων.

Με βάση λοιπόν τα στοιχεία των πινάκων που έχουμε δημιουργήσει μπορούμε να κατασκευάσουμε σχετικά γραφήματα που θα οπτικοποιήσουν τα στοιχεία μας. Τα γραφήματα αυτά μπορούν να έχουν τη μορφή διαγράμματος ή γραφικής παράστασης. Το διάγραμμα παρουσιάζει την εξέλιξη των στοιχείων με τη μορφή μιας συνεχούς τεθλασμένης γραμμής ή με τη μορφή ακίδων. Το γράφημα έχει ποικίλες εκφράσεις και μπορεί να έχει τη μορφή ράβδων, στηλών, πίτας, να είναι τρισδιάστατο ή να αποτελεί συνδυασμό με ένα διάγραμμα.

1.3 Παραδείγματα

Παράδειγμα 1

Σας δίνονται οι παρακάτω στατιστικές μεταβλητές. Εξηγήστε το είδος κάθε μίας:

α) Η εταιρία «Κτηματικές Επενδύσεις ΑΕ» διαθέτει 240 άτομα προσωπικό. Το προσωπικό της εταιρίας διακρίνεται με βάση τη θέση εργασίας ως εξής: 70 τοπογράφοι, 70 πολιτικοί μηχανικοί, 30 δικηγόροι, 20 πτυχιούχοι της πληροφορικής και 50 άτομα διοικητικό προσωπικό.

β) Η γενική αξιολόγηση που έλαβαν οι υπάλληλοι ενός υπουργείου κατά τη διαδικασία επιλογής νέου Γενικού Διευθυντή.

γ) Το επίπεδο μόρφωσης κατά φύλο των υπαλλήλων ενός φορέα.

δ) Το ωρομίσθιο των 400 εκτάκτων υπαλλήλων ενός δήμου.

ε) Ο αριθμός των παιδιών των οικογενειών που μένουν στο 14ο δημοτικό διαμέρισμα του δήμου Αθηναίων.

στ) Το ΑΕΠ των χωρών-μελών του ΟΟΣΑ.

Απάντηση:

α) Η στατιστική μεταβλητή που περιγράφεται είναι το επάγγελμα των υπαλλήλων της εταιρίας. Είναι φανερό ότι η μεταβλητή αυτή δεν μπορεί να περιγραφεί παρά μόνο λεκτικά. Πρόκειται δηλαδή για μια ποιοτική μεταβλητή. Μία πιθανή επεξεργασία που μπορούμε να κάνουμε είναι να δημιουργήσουμε ένα πίνακα με τα επαγγέλματα και να αντιστοιχίσουμε σε αυτά τον αριθμό των ατόμων που ανήκουν στο συγκεκριμένο επάγγελμα.

β) Στην περίπτωση αυτή η στατιστική μεταβλητή γενική αξιολόγηση των υπαλλήλων είναι κατ' αρχήν ποιοτική. Αυτό ισχύει δεδομένου ότι η αξιολόγηση έγινε με την απόδοση χαρακτηριστικών όπως άριστος, πολύ καλός, καλός, μέτριος κ.λπ. Έτσι μπορεί να υπάρξει όχι μόνο ομαδοποίηση αλλά και διάταξη ή ιεράρχηση των δεδομένων με κριτήριο την καλύτερη ή χειρότερη αξιολόγησή τους. Εάν όμως η αξιολόγηση έγινε στη βάση μιας κλίμακας π.χ. από το 0 έως το 100 (μόνο ακέραιοι βαθμοί) τότε έχουμε μια ποσοτική ασυνεχή μεταβλητή. Η μεταβλητή αυτή μπορεί να παρουσιαστεί με πλήρη κατάταξη των υπαλλήλων από τον καλύτερο προς τον χειρότερο.

γ) Στην περίπτωση αυτή έχουμε μια διπλή μεταβλητή. Ουσιαστικά πρόκειται για δύο μεταβλητές. Το επίπεδο μόρφωσης και το φύλο των εργαζομένων του φορέα.

- Η πρώτη μεταβλητή είναι καθαρά ποιοτική.
- Η δεύτερη μεταβλητή είναι ποιοτική και μπορεί να ταξινομηθεί σε δύο κατηγορίες άνδρας, γυναίκα.

δ) Εδώ πρόκειται για ποσοτική συνεχή μεταβλητή δεδομένου ότι το ωρομίσθιο μπορεί να λάβει δεκαδικές τιμές.

ε) Η μεταβλητή αυτή είναι ποσοτική ασυνεχής μεταβλητή αφού μπορεί να λάβει μόνο ακέραιες τιμές π.χ. μια οικογένεια μπορεί να έχει 1, 2, 3 παιδιά αλλά όχι 1,35.

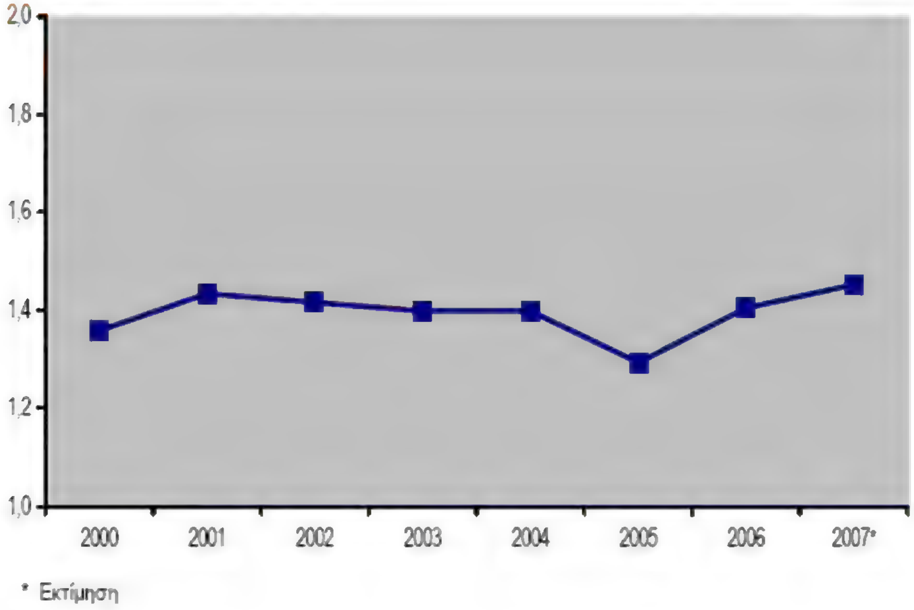
στ) Το ΑΕΠ είναι μια συνεχής ποσοτική μεταβλητή.

Παράδειγμα 2

Παρακάτω δίνονται μερικά παραδείγματα μορφών γραφημάτων που μπορούμε να χρησιμοποιούμε κατά την παρουσίαση των στατιστικών μας.

Το σχήμα 1.1 παρουσιάζει την περίπτωση της γραφικής παράστασης του λόγου των έμμεσων/άμεσους φόρους για την Ελλάδα.

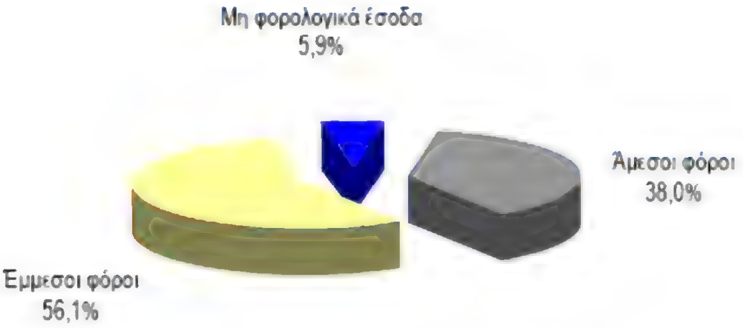
Σχήμα 1.1



Πηγή : ΥΠ.ΟΙΚ.Ο

Το σχήμα 1.2 παρουσιάζει ένα διάγραμμα με μορφή πίτας που δείχνει την ποσοστιαία σύνθεση των εσόδων του τακτικού προϋπολογισμού της Ελλάδας για το 2008.

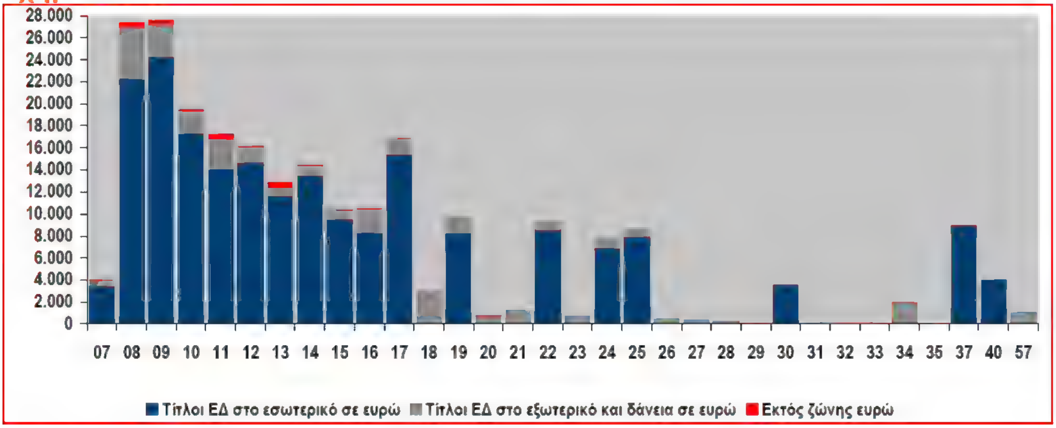
Σχήμα 1.2



Πηγή : ΥΠ.ΟΙΚ.Ο

Το σχήμα 1.3 παρουσιάζει ένα διάγραμμα με τη μορφή ραβδογράμματος που δείχνει το χρονοδιάγραμμα λήξης χρέους της κεντρικής κυβέρνησης στις 31.12.2007 για την Ελλάδα.

Σχήμα 1.3



Πηγή: ΥΠ.ΟΙΚ.Ο

2. Εκπαιδευτική Ενότητα

- **Το SPSS και το περιβάλλον εργασίας του**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να αναγνωρίζει τα βασικά μέρη του SPSS.
- Να αναφέρει τις δυνατότητες του SPSS.
- Να εξοικειωθεί με το περιβάλλον εργασίας του SPSS.
- Να αναγνωρίζει και να εφαρμόζει κωδικοποίηση-εισαγωγή δεδομένων.

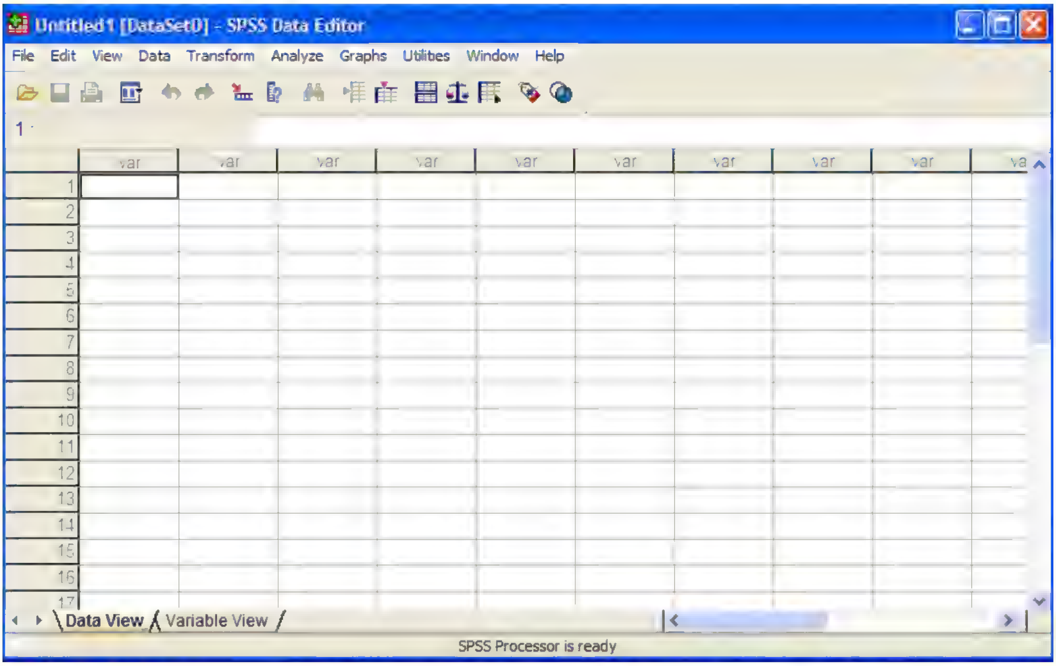
ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- SPSS Data Editor
- Διαδικασία Frequencies
- Διαδικασία Descriptives
- Κλίμακες Κατάταξης
- Πίνακες συχνοτήτων
- SPSS Output Navigator
- Παράθυρο Εντολών στο SPSS

2.1. Το SPSS και το περιβάλλον εργασίας του

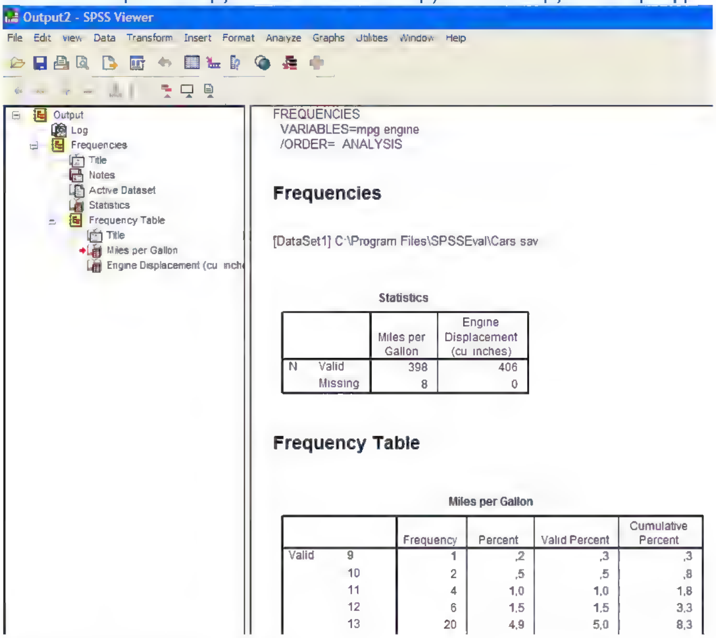
2.1.1 Στατιστικά πακέτα και πεδία εφαρμογής τους - Το πρόγραμμα SPSS και οι δυνατότητές του

Το SPSS είναι ένα από τα πιο εξελιγμένα υπολογιστικά προγράμματα Στατιστικής που προσπαθεί να καλύψει το σύνολο των γνωστότερων στατιστικών τεχνικών. Ιδιαίτερα στις εκδόσεις του σε γραφικό περιβάλλον, δίνει τη δυνατότητα γραφικής επεξεργασίας και αναδραστικής λειτουργίας με πολύ μεγάλη επιτυχία. Αντίστοιχα, υπάρχουν και άλλα στατιστικά πακέτα όπως το MINITAB, S-plus, SAS, STATA, R. Στην πλειοψηφία τους τα παραπάνω στατιστικά πακέτα με πιο γνωστά το S-plus και SAS είναι υπολογιστικά προγράμματα Στατιστικής με μεγάλη δυνατότητα προγραμματισμού διαδικασιών. Με τον τρόπο αυτό, προσφέρει στον έμπειρο χρήστη υψηλούς βαθμούς ελευθερίας στο χειρισμό συνόλων δεδομένων.
Τα βασικά μέρη από τα οποία αποτελείται το SPSS είναι το SPSS Data Editor, το Data View – Variable View και το Output SPSS, όπως παρουσιάζονται στις εικόνες 2.1-2.2.



Εικόνα 2.1: Η αρχική οθόνη του SPSS- SPSS Data Editor, Data View – Variable View.

Ταυτόχρονα με το SPSS Data editor, αν και όχι φανερά, το λογισμικό ενεργοποιεί και μια δεύτερη εφαρμογή, την SPSS Output Navigator (εικόνα 2.2). Σε αυτή το SPSS καταγράφει τα αποτελέσματα της κάθε στατιστικής ανάλυσης που πραγματώνει.



Εικόνα 2.2: Output1- SPSS Viewer

Το παράθυρο αποτελεσμάτων χωρίζεται σε δύο τμήματα: το ένα αποτελεί ένα πίνακα περιεχομένων για όλες τις δουλειές που πραγματοποιήθηκαν από το λογισμικό, ενώ το άλλο εμφανίζει αυτά καθαυτά τα αποτελέσματα. Επιλέγοντας (από τον πίνακα περιεχομένων) εκείνα που μας ενδιαφέρουν κάθε φορά, αυτόματα έχουμε τα αντίστοιχα αποτελέσματα στο διπλανό παράθυρο.

Το SPSS ακολουθεί τη μορφή των προγραμμάτων με συνεχείς επιλογές (menu driven- εικόνα 2.3). Πράγματι στην κύρια γραμμή εργαλείων εμφανίζονται 12 δυνατότητες και στο Data Editor και στο Output:

Εικόνα 2.3: Επιλογές (menu driven)

>> File: Χρησιμοποιούμε την επιλογή αυτή για να δημιουργήσουμε ένα καινούριο SPSS αρχείο, να διαβάσουμε ένα υπάρχον, ή να διαβάσουμε αρχεία δεδομένων που δημιουργήθηκαν με άλλα προγράμματα (EXCEL, LOTUS, dBASE, SPSS/PC+, κ.λπ).
>> Edit: Με την επιλογή αυτή μπορούμε να τροποποιήσουμε ή να αντιγράψουμε τμήματα του αρχείου δεδομένων.

>> **View:** Με τη βοήθεια της επιλογής αυτής μεταβάλλουμε τον αριθμό των πληροφοριών που υπάρχουν στο περιβάλλον εργασίας μας.

>> **Data:** Η επιλογή αυτή χρησιμοποιείται για να πραγματοποιήσουμε καθολικές αλλαγές στα δεδομένα, π.χ. ορισμό νέων μεταβλητών, συνένωση αρχείων δεδομένων, μετατόπιση μεταβλητών και περιπτώσεων, δημιουργία υποσύνολων δεδομένων κ.λπ.

>> **Transform:** Χρησιμοποιούμε την επιλογή αυτή για να πραγματοποιήσουμε αλλαγές σε κάποιες από τις μεταβλητές, όπως επανακωδικοποίηση τιμών, πράξεις μεταξύ των μεταβλητών, μετασχηματισμούς υπό συνθήκη, δημιουργία νέων μεταβλητών κ.λπ.

>> **Analyze:** Η επιλογή για την πραγμάτωση της στατιστικής ανάλυσης που επιθυμούμε.

>> **Graphs:** Χρησιμοποιούμε την επιλογή αυτή για να δημιουργήσουμε γραφικές παραστάσεις για τα δεδομένα μας, όπως ραβδογράμματα, ιστογράμματα, διαγράμματα διασποράς. Σε όλα τα γραφήματα μπορούμε να επέμβουμε με τη βοήθεια του «επεξεργαστή γραφημάτων» (Chart Editor) που είναι ενσωματωμένος στο SPSS.

>> **Utilities:** Η επιλογή αυτή μας δίνει τη δυνατότητα να δημιουργήσουμε γραφικά έναν πίνακα περιεχομένων για τις μεταβλητές του προβλήματος, να δούμε τις πληροφορίες για την καθεμία από αυτές ξεχωριστά και να δημιουργήσουμε/χρησιμοποιήσουμε μόνον μερικές από αυτές στη συγκεκριμένη στατιστική ανάλυση.

>> **Window:** Από την επιλογή αυτή μπορούμε να πληροφορηθούμε τα σχετικά με τα προγράμματα αρχεία (δεδομένων, αποτελεσμάτων, γραφημάτων, εντολών) που έχουμε ενεργοποιήσει κατά τη διάρκεια της στατιστικής ανάλυσης που πραγματοποιούμε, καθώς επίσης και να μετακινηθούμε μεταξύ τους.

>> **Help:** Χρησιμοποιούμε την επιλογή αυτή για να αποκτήσουμε πρόσβαση στην άμεση (on line) βοήθεια του λογισμικού.

2.1.2 Περιγραφική Στατιστική – Στατιστική παρουσίαση και εξέταση δεδομένων με τη χρήση του SPSS – Πίνακες Συχνοτήτων

Το πρώτο βήμα, στη μελέτη ενός συνόλου δεδομένων (δείγματος) είναι η παρουσίαση και ανάλυση των τιμών για τις τιμές που περιλαμβάνονται σε αυτό. Στην παράγραφο αυτή πρόκειται να αναλυθούν οι τεχνικές οι οποίες χρησιμοποιούνται για την παρουσίαση των συγκεντρωτικών στοιχείων της κάθε μεταβλητής χωριστά (Περιγραφική Στατιστική) και οδηγούν στην τοποθέτηση αρχικών υποθέσεων για τον πληθυσμό από τον οποίο προέρχονται (Στατιστική Συμπερασματολογία).

Η επιλογή της σωστής στατιστικής τεχνικής για την περιγραφή μιας μεταβλητής εξαρτάται αποκλειστικά από το χαρακτήρα της, τη διάκρισή της δηλαδή σε ποιοτική ή ποσοτική. Υπάρχει σαφής διαφοροποίηση των εργαλείων που είναι διαθέσιμα στην κάθε περίπτωση.

Ο όρος **δεδομένα (data)** αναφέρεται σε μετρήσεις ή παρατηρήσεις που προέρχονται από ένα πείραμα ή μια δειγματοληπτική έρευνα. Τα δεδομένα μπορεί να είναι ή **ποσοτικά (αριθμητικά)** ή **ποιοτικά (κατηγορικά)**. Συνοπτικά μπορούμε να πούμε τα εξής:

Οι τέσσερις χρησιμοποιούμενες κλίμακες μέτρησης είναι από την «ασθενέστερη» στην «ισχυρότερη», η ονομαστική κλίμακα (nominal scale), η διατεταγμένη κλίμακα (ordinal scale), η κλίμακα διαστήματος (interval scale) και η κλίμακα λόγου (ratio scale).

••• Η **ονομαστική κλίμακα (nominal scale)** κάνει χρήση αριθμητικών τιμών μόνο ως μέσων διαχωρισμού των ιδιοτήτων ή των στοιχείων σε διάφορες κατηγορίες. Οι τιμές δηλαδή είναι τα αυθαίρετα ονόματα των δυνατών κατηγοριών.

••• Η **διατεταγμένη κλίμακα (ordinal scale)** είναι μια ονομαστική κλίμακα στην οποία συγκρίσεις της μορφής "μεγαλύτερη", "μικρότερη", "ίση" μεταξύ των τιμών έχουν νόημα.

••• Η **κλίμακα διαστήματος (interval scale)** είναι μια διατεταγμένη κλίμακα στην οποία πέρα από τη σχετική διάταξη των μετρήσεων, έχει έννοια και το μέγεθος του διαστήματος των δύο μετρήσεων (δηλαδή το μέγεθος της διαφοράς με την έννοια της αφαίρεσης των σχετικών τιμών). Το μηδέν ορίζεται αυθαίρετα στην κλίμακα αυτή, όπως και η μονάδα (μοναδιαία απόσταση).

••• Η **κλίμακα λόγου (ratio scale)** είναι μια κλίμακα διαστήματος στην οποία, πέρα από τη διάταξη και το μέγεθος του διαστήματος μεταξύ δύο τιμών, έχει έννοια η σύγκριση δύο τιμών μέσω του λόγου τους. Το μηδέν ορίζεται στην κλίμακα αυτή, δηλαδή υπάρχει φυσική μέτρηση που ονομάζεται μηδέν. Η μονάδα ορίζεται αυθαίρετα.

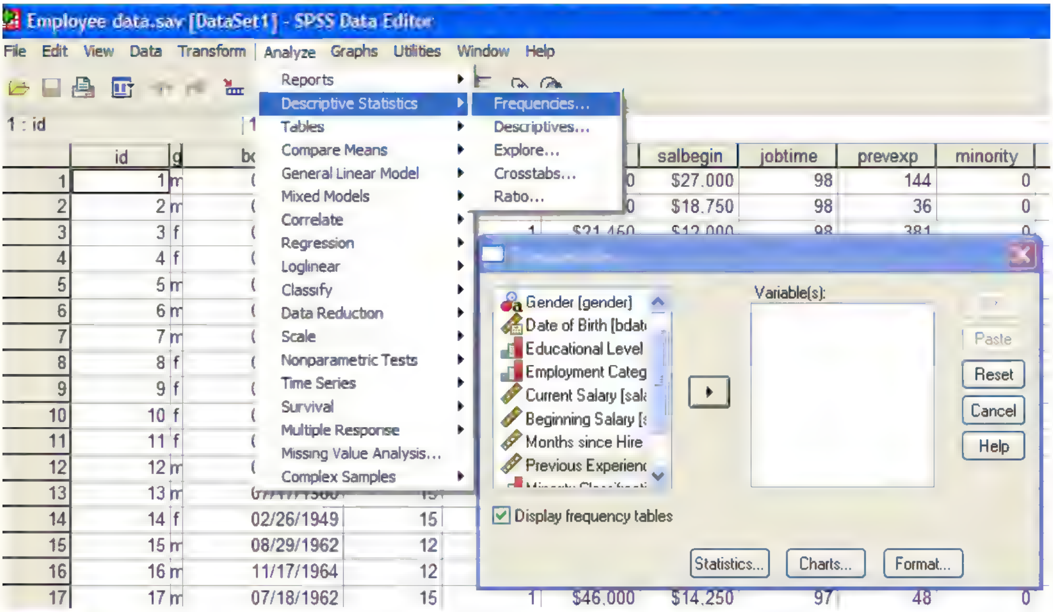
Συνήθως, τα πρωτογενή δεδομένα που έχουμε σε μια μελέτη είναι τεράστια σε μέγεθος και έχουν ακατάστατη μορφή με αποτέλεσμα να μην μπορούμε να διακρίνουμε τις πληροφορίες που περιέχουν. Οι μέθοδοι της Περιγραφικής Στατιστικής αποτελούν το επιστημονικό εργαλείο για τη σύνοψη, ταξινόμηση και παρουσίασή τους σε εύληπτη μορφή. Τρεις διαφορετικές τεχνικές μπορούν να χρησιμοποιηθούν: **οι πίνακες συχνοτήτων, οι γραφικές παραστάσεις και τα στατιστικά μέτρα.** Το SPSS ενσωματώνει διάφορες διαδικασίες για την πραγμάτωση Περιγραφικής Στατιστικής στις παρατηρήσεις/τιμές μιας μεταβλητής. Η επιλογή της κατάλληλης μεταξύ αυτών εξαρτάται αποκλειστικά από το χαρακτήρα των δεδομένων, τη διάκρισή τους δηλαδή σε ποιοτικά και ποσοτικά.

Ποιοτικά χαρακτηριστικά

Οι μέθοδοι σύνοψης και παρουσίασης ποιοτικών δεδομένων περιορίζονται στους πίνακες συχνοτήτων και τις γραφικές παραστάσεις. Με τη διαδικασία "Frequencies" μπορούμε να πετύχουμε την άμεση κατασκευή τους. Η διαδικασία αυτή μπορεί να χρησιμοποιηθεί και στην περίπτωση των ποσοτικών μεταβλητών.

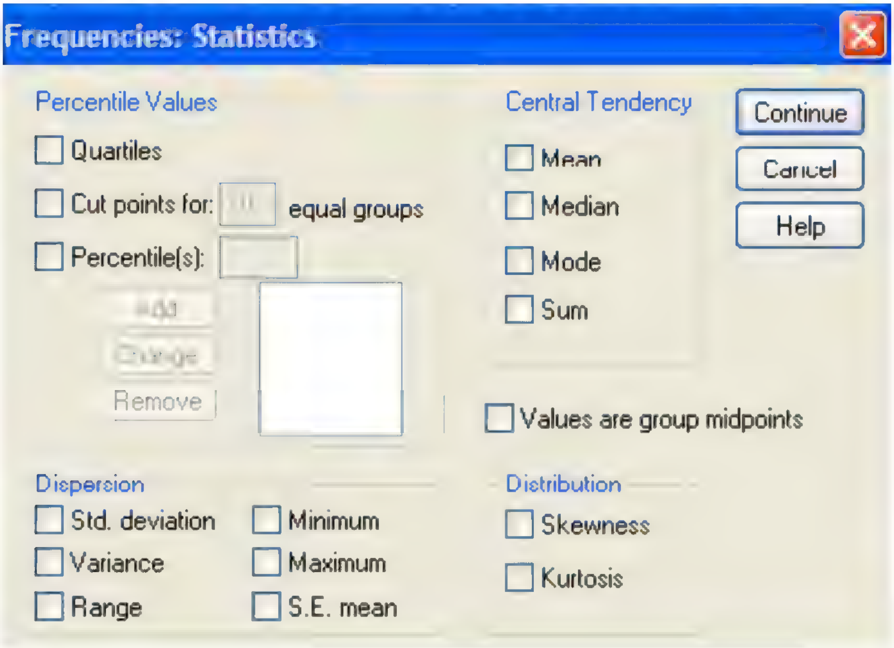
- Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγουμε:

Analyze => Descriptive Statistics => Frequencies



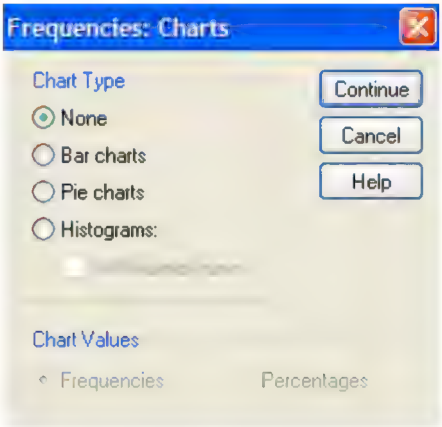
Εικόνα 2.4: Πλαίσιο διαλόγου του SPSS, για την εφαρμογή της διαδικασίας Frequencies

Διαλέγουμε τη μεταβλητή (ή τις μεταβλητές) που θέλουμε να αναλύσουμε και τις μετακινούμε στο παράθυρο **Variables(s)**. Εξ ορισμού η διαδικασία θα κατασκευάσει μόνον τον αντίστοιχο πίνακα συχνοτήτων. Το λογισμικό δε θα υπολογίσει κανένα από τα στατιστικά μέτρα (εξάλλου, μόνον η επικρατούσα τιμή έχει νόημα). Η επιλογή **"Statistics"** δίνει πρόσβαση στα πιο συνηθισμένα από αυτά.



Εικόνα 2.5: Πλαίσιο διαλόγου του SPSS, για τα στατιστικά μέτρα της διαδικασίας Frequencies

Καμία γραφική παράσταση δε θα κατασκευαστεί. Αν επιθυμούμε κάποιο γράφημα θα πρέπει να ενεργοποιήσουμε την επιλογή **"Charts"**, που φαίνεται στην εικόνα 2.4.

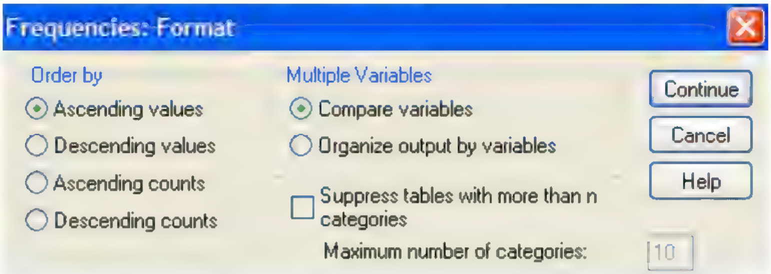


Εικόνα 2.6: Πλαίσιο διαλόγου του SPSS, για τα γραφήματα της διαδικασίας Frequencies

◇ **Chart Type:** Μπορούμε να κατασκευάσουμε ραβδογράμματα, κυκλικά διαγράμματα και ιστογράμματα. Φυσικά τα τελευταία αφορούν την περίπτωση των ποσοτικών μεταβλητών. Επιλέγοντας την εμφάνιση της κανονικής καμπύλης (Histogram With normal curve) βοηθούμαστε σε μια γραφική εκτίμηση της ύπαρξης ή προσέγγισης των δεδομένων μας από την κανονική κατανομή.

◇ **Chart Value:** Καθορίζουμε αν στον κατακόρυφο άξονα των υπό κατασκευή ραβδογραμμάτων (ή ιστογραμμάτων) θα εμφανίζονται οι απόλυτες (**Frequencies**) ή οι σχετικές τιμές (**Percentages**).

Καθορίζουμε τον τρόπο εμφάνισης των κατηγοριών του πίνακα συχνοτήτων από την επιλογή Format, που παρουσιάζεται στην εικόνα 2.4. Ο πίνακας συχνοτήτων μπορεί να ταξινομηθεί (εικόνα 2.7) ως προς τη σειρά εμφάνισης των διαφορετικών κατηγοριών στα δεδομένα ως προς τη συχνότητα εμφάνισης τους (και μάλιστα κατά αύξουσα - Ascending values ή φθίνουσα –Descending values σειρά).



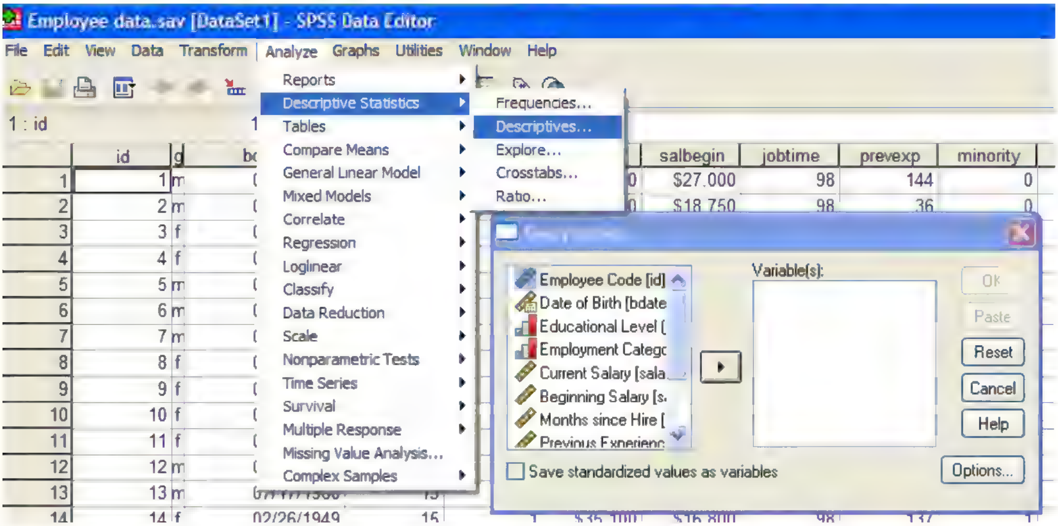
Εικόνα 2.7: Πλαίσιο διαλόγου του SPSS, για τον καθορισμό του τρόπου εμφάνισης των κατηγοριών του πίνακα συχνοτήτων, της διαδικασίας Frequencies.

Ποσοτικά χαρακτηριστικά

Η διαδικασία **"Descriptives"** βοηθάει στον υπολογισμό των στατιστικών μέτρων μιας ή περισσοτέρων ποσοτικών μεταβλητών.

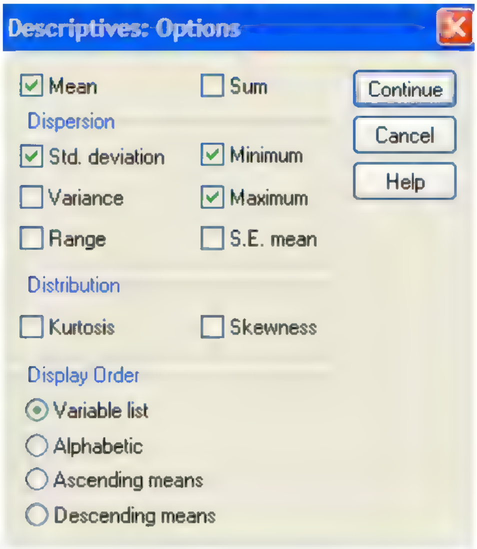
√ Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγουμε:

Analyze => Descriptive Statistics => Descriptives



Εικόνα 2.8: Πλαίσιο διαλόγου του SPSS, για την εφαρμογή της διαδικασίας Descriptives

Διαλέγουμε τη μεταβλητή (ή τις μεταβλητές) που θέλουμε να περιγράψουμε και τη μετακινούμε στο παράθυρο **Variable(s)** (εικόνα 2.8). Εξ ορισμού η διαδικασία θα προχωρήσει στον υπολογισμό των στατιστικών μέτρων: μέση τιμή, τυπική απόκλιση, μικρότερη και μεγαλύτερη τιμή. Η επιλογή "Options" (εικόνα 2.9) μας δίνει τη δυνατότητα να βρούμε τις τιμές και άλλων μέτρων και να καθορίσουμε τη σειρά εμφάνισής τους.



Εικόνα 2.9: Πλαίσιο διαλόγου του SPSS, για τα στατιστικά μέτρα της διαδικασίας Descriptives

Τέλος μπορούμε να ζητήσουμε τον υπολογισμό και στη συνέχεια την αποθήκευσή των τυποποιημένων τιμών (standardized values) για τις μεταβλητές του καταλόγου Variable(s) (εικόνα 2.8). Η διαδικασία "**Descriptives**" είναι μια φτωχή ως προς τις δυνατότητες διαδικασία η οποία προορίζεται για την ταυτόχρονη-γρήγορη παρουσίαση (στον ίδιο πίνακα) των στατιστικών μέτρων δύο ή περισσότερων ποσοτικών μεταβλητών. Όπου το ενδιαφέρον μας μπορεί να εξειδικευτεί, θα πρέπει να αντικαθίσταντα από την **"Frequencies"**.

2.1.3 Το αρχείο αποτελεσμάτων του SPSS

Το αποτέλεσμα της τυχαίας διαδικασίας που ολοκλήρωσε το SPSS, από τη δημιουργία μιας απλής γραφικής παράστασης μέχρι μια πολύπλοκη στατιστική ανάλυση, εμφανίζεται στην εφαρμογή **"SPSS Output Navigator"**. Στην εφαρμογή αυτή καταγράφονται τα αποτελέσματα όλων των διαδικασιών (στατιστικών αναλύσεων, γραφικών παραστάσεων, μετατροπών, κ.λπ) που πραγματοποιήθηκαν σε ένα σύνολο δεδομένων, και μάλιστα με τη σειρά που εκτελέστηκαν.

Το παράθυρο αποτελεσμάτων χωρίζεται σε δύο τμήματα: το αριστερό (Outline pane) εμφανίζει έναν πίνακα περιεχομένων για όλες τις δουλειές που πραγματοποιήθηκαν από το λογισμικό, ενώ το δεξί (Display Pane) εμφανίζει αυτά καθαυτά τα αποτελέσματα (εικόνα 2.2).

Outline pane (πίνακας περιεχομένων των αποτελεσμάτων)

Παρατηρήστε ότι η κάθε διαδικασία έχει αναλυθεί σε επιμέρους συστατικά στοιχεία. Στο καθένα από αυτά, το λογισμικό αντιστοιχεί το εικονίδιο ενός βιβλίου, συνήθως ανοιχτού (εικόνα 2.2). Επιλέγοντας (από το αριστερό τμήμα) εκείνο ακριβώς το σημείο που μας ενδιαφέρει, αυτόματα παίρνουμε δίπλα τα αντίστοιχα αποτελέσματα.

Ένα κλειστό βιβλίο, υποδηλώνει την ύπαρξη αποτελεσμάτων τα οποία καλύφθηκαν και συνεπώς δεν εμφανίζονται. Μπορούμε όμως (με διπλό κλικ στο εικονίδιο) να αντιστρέψουμε αυτή την κατάσταση. Φυσικά ισχύει και το αντίστροφο: διπλό κλικ με το εικονίδιο ενός ανοιχτού βιβλίου θα το μετατρέψει σε κλειστό και τα αντίστοιχα αποτελέσματα θα πάψουν να εμφανίζονται στο δεξί τμήμα.

Επιπρόσθετα, μπορούμε να καλύψουμε την εμφάνιση όλων των αποτελεσμάτων μιας διαδικασίας που πραγματοποιήθηκε κι όχι μόνον τμημάτων της. Αρκεί να μετατρέψουμε το σύμβολο του μείον –με απλό κλικ στο τετραγωνάκι που περιέχεται– σε συν.

Display pane (τα αποτελέσματα)

Το δεξί τμήμα του παραθύρου των αποτελεσμάτων εμφανίζει μόνον όσα από τα αποτελέσματα χωρούν σε αυτό. Για να δούμε και τα υπόλοιπα, ή χρησιμοποιούμε το αριστερό μέρος (επιλογή από τον κατάλογο περιεχομένων), ή κινούμαστε μέσα σε αυτό. Τα αποτελέσματα που περιέχονται εδώ αντιστοιχούν στα συστατικά στοιχεία που εμφανίζονται δίπλα και χαρακτηρίζονται σαν αντικείμενα: πίνακες αριθμών, γραφικές παραστάσεις, κείμενα. Όλα τα αντικείμενα μπορούμε να τα τροποποιήσουμε ενεργοποιώντας τον αντίστοιχο επεξεργαστή (με διπλό κλικ πάνω στο αντικείμενο) που έχει ενσωματωμένο το λογισμικό.

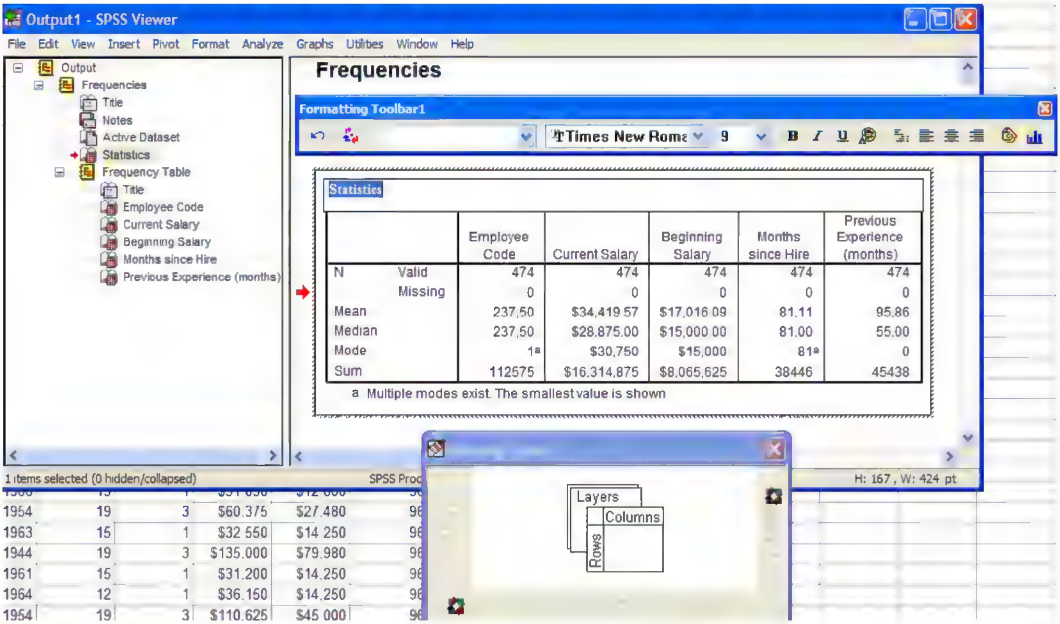
Πίνακες Αριθμών *(Pivot Tables)*

Το SPSS εμφανίζει όλα τα αποτελέσματα μιας στατιστικής ανάλυσης σε μορφή ενός πίνακα τον οποίο και μπορούμε να επεξεργαστούμε:
---- Στο αριστερό μέρος επιλέγουμε το εικονίδιο που αντιστοιχεί στην περιγραφική ανάλυση της μεταβλητής που θέλουμε να επεξεργαστούμε. Ο πίνακας εμφανίζεται αμέσως στο δεξί μέρος του παραθύρου αποτελεσμάτων.

---- Μετακινούμε τον δείκτη του ποντικιού πάνω στον πίνακα και με διπλό κλικ ειδοποιούμε το λογισμικό ότι θέλουμε να επέμβουμε στην εμφάνισή του. Ο πίνακας περιβάλλεται τώρα από μια γραμμή σκίασης που υποδεικνύει την ενεργοποίηση του και συνεπώς και τη δυνατότητα επέμβασης σε αυτόν. Επιπλέον, η βασική γραμμή των επιλογών έχει αλλάξει (έχει εμφανιστεί και η επιλογή "Pivoting Trays 1").

Στον πίνακα που επιλέξαμε έχουμε τη δυνατότητα να τροποποιήσουμε τα πάντα. Αν για παράδειγμα δεν μας αρέσει ο τίτλος "Statistics" που το λογισμικό έδωσε, αφού πρώτα τον επιλέξουμε, αρκεί να κάνουμε διπλό κλικ επάνω του. Φυσικά μπορούμε αλλάξουμε τη γραμματοσειρά, ή να κάνουμε το κείμενο έντονο, πλάγιο, μεγαλύτερο κ.λπ. Επιλέγουμε τον τίτλο και στη συνέχεια από τη βασική ράβδο των προτιμήσεων δίνουμε τη μορφή που εμείς θέλουμε.

Οι αλλαγές μορφοποίησης που μπορούμε να κάνουμε σε έναν πίνακα αριθμών τον οποίο έχουμε ήδη επιλέξει, είναι προσιτές από τη βασική ράβδο προτιμήσεων μέσω των διαδοχικών επιλογών. Όλα τα παραπάνω παρουσιάζονται στην εικόνα 2.10.



Εικόνα 2.10: Ενεργοποίηση του επεξεργαστή πινάκων-γραφημάτων (pivot table)

Εκτύπωση Αποτελεσμάτων

Για να εκτυπώσουμε το περιεχόμενο του παραθύρου αποτελεσμάτων αρκεί να επιλέξουμε διαδοχικά:
File => Print => OK

από τη βασική ράβδο προτιμήσεων του λογισμικού στο παράθυρο των αποτελεσμάτων. Επιπλέον υπάρχει η δυνατότητα να τυπώσουμε μέρος μόνο των αποτελεσμάτων που έχουμε. Στο αριστερό μέρος επιλέγουμε τα εικονίδια των αποτελεσμάτων που μας ενδιαφέρουν (συνδυασμός ποντικιού με το πλήκτρο Ctrl). Από τη ράβδο επιλέγουμε διαδοχικά:
File => Print => OK

Για την καλύτερη εμφάνιση των αποτελεσμάτων στην εκτύπωση, χρήσιμη είναι η διαδικασία της προεπισκόπησης με τη βοήθεια του κουμπιού. Αν διαπιστώσουμε ότι σε κάποια θέση πρέπει να αρχίσουμε καινούρια σελίδα, επιστρέφουμε στον πίνακα αποτελεσμάτων, επιλέγουμε το συγκεκριμένο πίνακα και ζητάμε αλλαγή σελίδας με τον ακόλουθο τρόπο:

Insert => Page Break

Αποθήκευση Αποτελεσμάτων

Για να αποθηκεύσουμε το περιεχόμενο του παραθύρου αποτελεσμάτων από τη βασική ράβδο (menu) διαδοχικά επιλέγουμε:
File => Save as

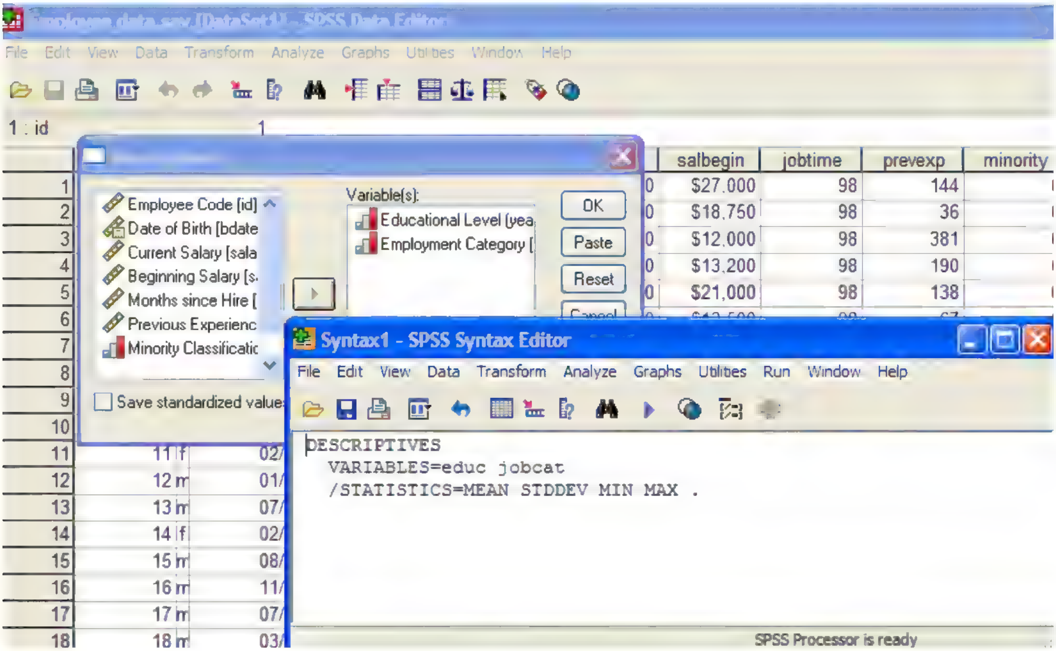
Και στη συνέχεια δίνουμε το όνομα που επιθυμούμε, Τα αρχεία αποτελεσμάτων του SPSS έχουν την κατάληξη **".spo".**

2.1.4 Το αρχείο εντολών του SPSS

Το SPSS for Windows, επιτρέποντας την πραγμάτωση της ανάλυσης με τη βοήθεια των διαδοχικών οθονών επιλογής (menu driven), κατάργησε, στις πιο πολλές τουλάχιστον περιπτώσεις, την ανάγκη δημιουργίας τέτοιων αρχείων. Είναι επιπλέον συχνά χρήσιμο να έχουμε αποθηκευμένες κάπου χωριστά τις εντολές που χρησιμοποιήθηκαν σε κάποια στατιστική ανάλυση, έτσι ώστε να μπορούμε να την επαναλάβουμε ύστερα από κάποιες αλλαγές των δεδομένων χωρίς να χρειάζεται να περάσουμε ξανά από τις οθόνες επιλογής. Ένα άλλο σημαντικό πλεονέκτημα των αρχείων αυτών, είναι ότι μπορούν να δημιουργηθούν με ένα οποιοδήποτε ASCHII κειμενογράφο σαν ένα κοινό αρχείο. Στη συνέχεια εισάγονται στο SPSS ως "αρχείο εντολών" και εκτελούνται.

Γενικότερα υπάρχουν τρεις τρόποι για να δημιουργήσουμε ένα "αρχείο εντολών":

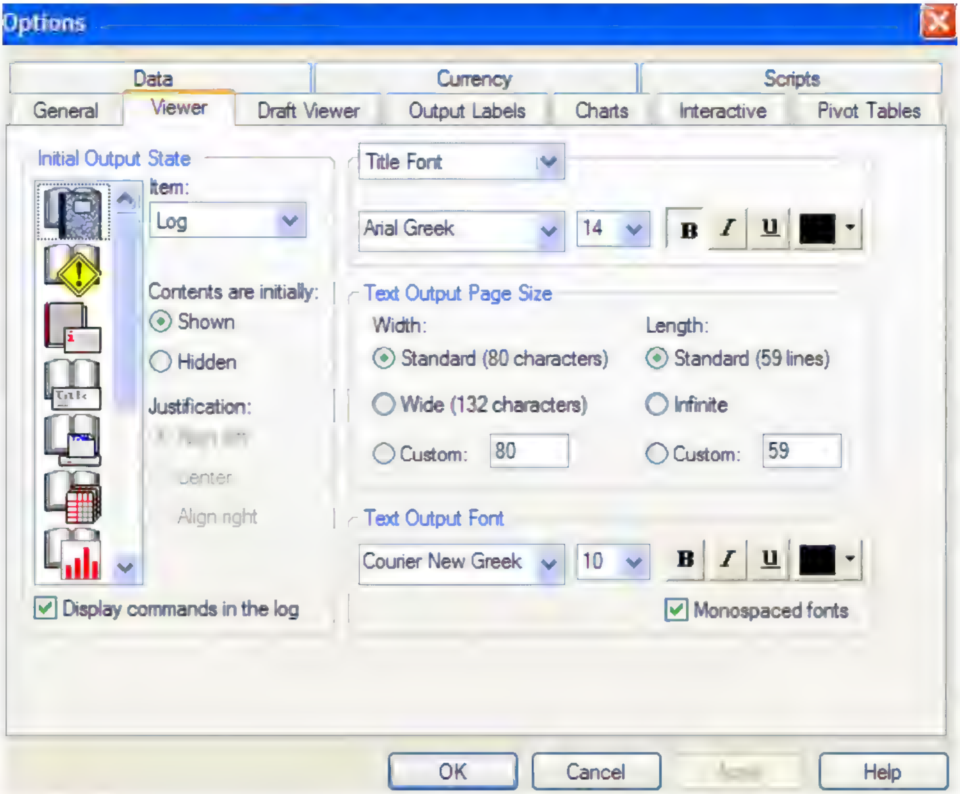
- Ο πιο εύκολος τρόπος είναι κατά τη διάρκεια των διαδοχικών επιλογών να χρησιμοποιούμε το πλήκτρο "Paste". Αυτό έχει σαν αποτέλεσμα να αντιγράφονται τα βήματα της επιχειρούμενης στατιστικής ανάλυσης διαδοχικά σε ένα αρχείο εντολών που ενεργοποιείται αυτόματα το λογισμικό του (εικόνα 2.11).



Εικόνα 2.11: Δημιουργία αρχείου εντολών με τη χρήση του πλήκτρου "*Paste*"

- Μπορούμε να χρησιμοποιήσουμε τα εικονίδια "log" τα οποία εμφανίζονται στο παράθυρο των αποτελεσμάτων, αν έχουμε επιλέξει:

Edit => Options => και στην καρτέλα Viewer επιλέξουμε Display Commands in the log

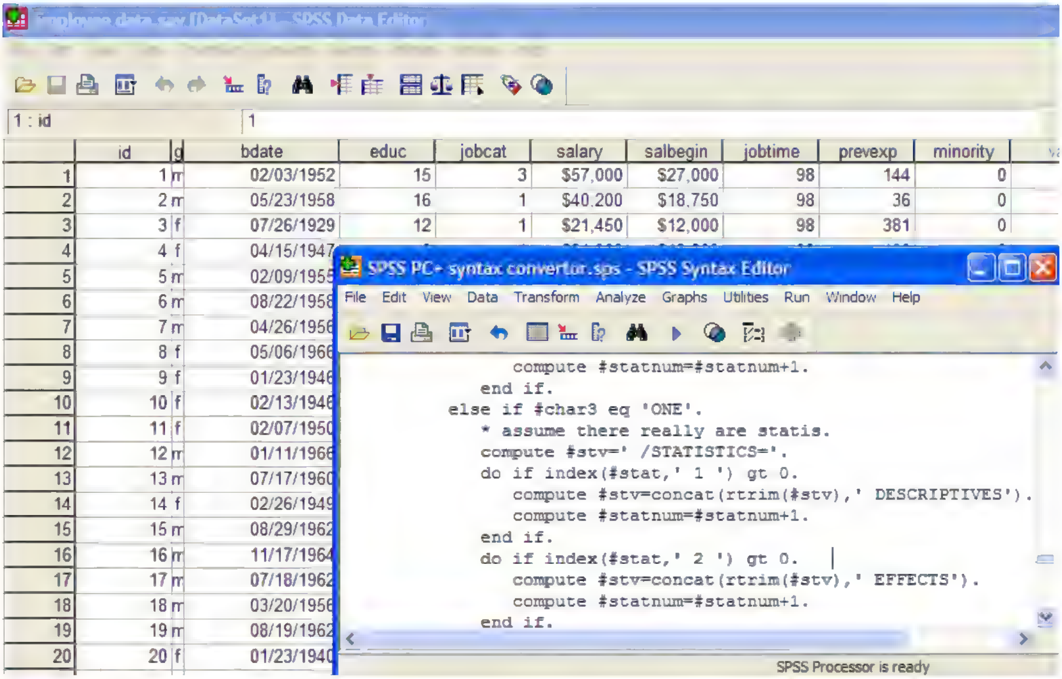


Εικόνα 2.12: Δημιουργία αρχείου εντολών με τη χρήση των εικονιδίων "log"

πριν εκτελέσουμε την επιχειρούμενη στατιστική ανάλυση. Η σύνταξη της κάθε εντολής εμφανίζεται στο παράθυρο αποτελεσμάτων ακριβώς πριν τη δημιουργία του.

• Μπορούμε να χρησιμοποιήσουμε το SPSS Syntax Editor, όπου το λογισμικό καταγράφει όλες τις εντολές που δίνονται κατά τη διάρκεια μιας στατιστικής ανάλυσης, ώστε με τη λήξη εκτέλεσης του προγράμματος να έχουμε ένα ξεχωριστό αρχείο. Για να αποκτήσουμε πρόσβαση σε αυτό επιλέγουμε διαδοχικά:

File Open => Syntax



Εικόνα 2.13: Δημιουργία αρχείου εντολών SPSS Syntax Editor

Στη συνέχεια, καθοδηγούμε το λογισμικό στον εντοπισμό του αντίστοιχου αρχείου. Το αποθηκεύουμε ως αρχείο εντολών ".sps" και επισημαίνουμε ότι με κάθε ξεκίνημα του SPSS το λογισμικό διαγράφει το υπάρχον αρχείο εντολών και δημιουργεί ένα καινούριο.

Εκτέλεση του περιεχόμενου ενός αρχείου εντολών.

Αφού διαβάσαμε ένα αρχείο εντολών μπορούμε με τη βοήθεια του menu "Run" να:

1. **All:** εκτελέσουμε όλες τις εντολές που περιέχονται σε αυτό.
2. **Selection:** εκτελέσουμε μόνο τις εντολές που περιέχοντα σε αυτό.
3. **Current:** εκτελέσουμε μόνο την εντολή στην οποία βρίσκεται ο δρομέας.
4. **To end:** εκτελέσουμε όλες τις εντολές από το σημείο που βρίσκεται ο δρομέας μέχρι το τέλος του αρχείου.

3. Εκπαιδευτική Ενότητα

- **Περιγραφική στατιστική Ι: πίνακες και διαγράμματα**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατασκευάζει στατιστικούς πίνακες μονομεταβλητών και πολυμεταβλητών πληθυσμών.
- Να κατανοεί τον τρόπο απεικόνισης των κατανομών συχνοτήτων και να κατασκευάζει πίνακες συνάφειας.
- Να απεικονίζει την πληροφορία που διαθέτει μέσα από Στατιστικά Διαγράμματα-Ραβδογράμματα-Υποδιαιρούμενα ακιδωτά-Κυκλικά-Χρονοδιαγράμματα-Διαγράμματα διασποράς-Ιστογράμματα-Πολυγωνικές γραμμές κ.ά.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Στατιστικός πληθυσμός
- Παρατήρηση
- Κατανομή Συχνότητας
- Ιστογράμματα Συχνοτήτων
- Πολυγωνική Γραμμή
- Αθροιστική κατανομή συχνοτήτων
- Σχετική κατανομή συχνότητας

Αναφέραμε στο προηγούμενο κεφάλαιο ότι η στατιστική ανάλυση αποτελεί μια διαδικασία που περιλαμβάνει πέντε συγκεκριμένα στάδια. Ήδη αναφέραμε τα τρία από αυτά. Προκειμένου όμως να προχωρήσουμε στο τέταρτο στάδιο, αυτό της ανάλυσης των στατιστικών στοιχείων και εξαγωγής στατιστικών συμπερασμάτων, θα πρέπει να αναφερθούμε σε κάποιες πρόσθετες έννοιες, αυτές του στατιστικού πληθυσμού, της παρατήρησης και της κατανομής συχνότητας.

Πολλοί συγχέουν την έννοια της παρουσίασης με αυτή της ανάλυσης των στοιχείων. Παρακάτω θα γίνει κατανοητό ότι η ανάλυση των στοιχείων ξεκινάει από τη στιγμή που δημιουργούμε νέα στοιχεία από τα ήδη υπάρχοντα. Τα νέα αυτά στοιχεία παρουσιάζονται και με τη βοήθεια πινάκων, διαγραμμάτων και γραφημάτων, αλλά έχουμε ήδη περάσει στο στάδιο της ανάλυσης, μιας και τα στοιχεία μας δεν είναι πλέον τα αρχικά ληφθέντα.

3.1 Η έννοια του στατιστικού πληθυσμού

Όπως αναφέρθηκε και στο κεφάλαιο 1 το πρώτο πράγμα που θα πρέπει να κάνουμε είναι να αποκτήσουμε μια σαφή εικόνα για αυτό που αποτελεί το στατιστικό πληθυσμό μας. Θυμηθείτε ότι η ποιότητα των στοιχείων μας είναι αυτή που θα μας βοηθήσει να εξαγάγουμε υψηλής ποιότητας συμπεράσματα. Εάν ο πληθυσμός μας δεν είναι πλήρως διασαφηνισμένος τότε η εξαγωγή του δείγματος που θα πραγματοποιήσουμε θα είναι λανθασμένη.

Ως πληθυσμό ορίζουμε ένα σύνολο μονάδων με κάποιο κοινό χαρακτηριστικό το οποίο αποτελεί για μας αντικείμενο διερεύνησης. Ένα παράδειγμα είναι το σύνολο των υπαλλήλων ενός υπουργείου που παρακολουθεί το σεμινάριο «Διοίκηση μέσω Στόχων» της Εθνικής Σχολής Δημόσιας Διοίκησης μέσα σε ένα έτος. Το σύνολο αυτών των υπαλλήλων είναι ένας στατιστικός πληθυσμός με κοινά χαρακτηριστικά το συγκεκριμένο σεμινάριο. Από τον πληθυσμό αυτό ορισμένες μονάδες είναι άνδρες και ορισμένες μονάδες γυναίκες. Άλλες μονάδες παρακολούθησαν με επιτυχία το σεμινάριο και άλλες όχι. Η στατιστική μας ανάλυση που περιλαμβάνει αυτή ακριβώς τη διάκριση με βάση κάποιο κοινό χαρακτηριστικό είναι μέρος της στατιστικής ανάλυσης.

Η έννοια του πληθυσμού όπως παρουσιάστηκε στο παραπάνω παράδειγμα είναι αυτή του μονοδιάστατου πληθυσμού. Αυτός διακρίνεται από τον πολυδιάστατο πληθυσμό που έχουμε στις περιπτώσεις που ο στατιστικός μας πληθυσμός διακρίνεται με βάση περισσότερα από ένα κοινά χαρακτηριστικά.

Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο το να αναλύσουμε ολόκληρο τον πληθυσμό είναι εκτός από μακρόχρονη και επίπονη διαδικασία και μια συχνά υψηλού κόστους επιλογή. Για το λόγο αυτό σχεδόν πάντα όταν ο πληθυσμός μας είναι πολυπληθής εφαρμόζουμε την τεχνική της δειγματοληψίας που αναλύεται σε επόμενο κεφάλαιο. Με την τεχνική αυτή αντί για τη μελέτη του συνολικού πληθυσμού επιλέγουμε ένα δείγμα. Η επιλογή του δείγματος αυτού γίνεται με τυχαίο τρόπο. Η εφαρμογή των κανόνων της δειγματοληψίας μας επιτρέπει να εξάγουμε ένα δείγμα που να αποτελεί μια μικρογραφία του συνολικού πληθυσμού. Επιδιώκουμε δηλαδή να έχει το δείγμα αυτό ή να τείνει να έχει τα ίδια χαρακτηριστικά με όλες τις μονάδες του πληθυσμού. Η επιλογή και η ανάλυση ενός δείγματος από ένα στατιστικό πληθυσμό περιορίζει σημαντικά το κόστος και αποτελεί ουσιαστικά το αντικείμενο και το στόχο της στατιστικής επιστήμης. Όποια συμπεράσματα εξάγουμε για το δείγμα μας μπορούμε στη συνέχεια να τα ανάγουμε στο σύνολο του πληθυσμού εξασφαλίζοντας, όπως θα δούμε, και συγκεκριμένα περιθώρια για τυχόν σφάλματα στην αναγωγή μας.

3.2 Η έννοια της παρατήρησης

Το σύνολο των παρατηρήσεων που επιλεγούμε να μελετήσομε αποτελεί μια μεταβλητή, ποσοτική ή ποιοτική, συνεχή ή ασυνεχή η οποία στη συνέχεια αποτελεί αντικείμενο της στατιστικής ανάλυσης.

Ο συνολικός πληθυσμός μας απαρτίζεται από μονάδες. Κάθε μονάδα του στατιστικού πληθυσμού αποκαλείται παρατήρηση. Η κάθε παρατήρηση διαθέτει δύο χαρακτηριστικά: είναι αυτοτελής και αυτόνομη. Το πρώτο χαρακτηριστικό είναι απόρροια του ότι η κάθε παρατήρηση από μόνη της παρουσιάζει όλα τα χαρακτηριστικά γνωρίσματα του πληθυσμού και το δεύτερο είναι αποτέλεσμα του ότι η κάθε παρατήρηση δεν επηρεάζεται από τις άλλες.

Σημειώνουμε εδώ ότι όλα τα παραπάνω ισχύουν και στην περίπτωση που μελετάμε ένα δείγμα αντί του συνολικού πληθυσμού. Με βάση τις δύο αυτές ιδιότητες η κάθε παρατήρηση αποτελεί αναπόσπαστο κομμάτι του πληθυσμού. Αυτό σημαίνει ότι εάν αντικατασταθεί με άλλη θα υπάρξει διαφορετικός συνδυασμός παρατηρήσεων. Το πρόβλημα που ενδέχεται να ανακύψει στην περίπτωση αυτή είναι να οδηγηθούμε σε άλλα συμπεράσματα αναφορικά με τη συμπεριφορά του πληθυσμού.

3.3 Η έννοια της κατανομής συχνότητας

Αφού έχουμε συγκεντρώσει τα στατιστικά μας στοιχεία και έχουμε πραγματοποιήσει μια πρώτη ταξινόμηση αυτών περνάμε στο τέταρτο στάδιο της στατιστικής ανάλυσης. Στο στάδιο αυτό χρησιμοποιούμε ειδικές κατατάξεις των στοιχείων μας που ονομάζονται κατανομές συχνοτήτων. Το ακριβές είδος και ο τρόπος κατασκευής των κατανομών εξαρτάται από τη συγκεκριμένη μεταβλητή που εξετάζουμε.

3.3.1 Η περίπτωση της ασυνεχούς μεταβλητής

Όταν η μεταβλητή που εξετάζουμε είναι ασυνεχής και μάλιστα με μικρό εύρος τιμών (π.χ. ο αριθμός των παιδιών κάθε υπαλλήλου ενός υπουργείου) ακολουθούμε την τεχνική της άθροισης των παρατηρήσεων ανά τιμή της μεταβλητής και παρουσιάζουμε τα στοιχεία σε ένα πίνακα απλής εισόδου. Η πρώτη στήλη του πίνακα περιλαμβάνει τις τιμές της μεταβλητής μας και η δεύτερη τον αριθμό των παρατηρήσεων που αναφέρονται σε κάθε τιμή της μεταβλητής. Αυτός ο αριθμός των παρατηρήσεων ανά τιμή της μεταβλητής ονομάζεται συχνότητα και η στήλη του πίνακα στήλη συχνοτήτων.

Εάν οι τιμές της ασυνεχούς μεταβλητής είναι πολλές (π.χ. η βαθμολογία σε ένα τεστ δεξιοτήτων με κλίμακα βαθμών από 0-100) μπορούμε αντί για τιμές στη στήλη των τιμών να τοποθετούμε μια τάξη τιμών (π.χ. βαθμολογία στο τεστ δεξιοτήτων από 0-40, από 41-80, και από 81-100).

Στον Πίνακα 3.1 παρουσιάζουμε στοιχεία βαθμολογίας του τεστ δεξιοτήτων από ένα δείγμα υπαλλήλων ενός υπουργείου. Η μεταβλητή Χ στον Πίνακα 3.1 παίρνει τις τιμές 0, 1, 2, ..., 99, 100 (βαθμολογία τεστ) και εμφανίζεται στην πρώτη στήλη με τις τιμές όμως ταξινομημένες σε τάξεις λόγω του μεγάλου εύρους τιμών που έχουμε. Η δεύτερη στήλη παρουσιάζει την κεντρική τιμή, μια εικονική τιμή της μεταβλητής ή αλλιώς μια αντιπροσωπευτική τιμή της συγκεκριμένης τάξης. Προσδιορίζεται από το άθροισμα και στη συνέχεια τη διαίρεση διά δύο των άκρων της συγκεκριμένης τάξης. Η τρίτη στήλη f_t παρουσιάζει τις αντίστοιχες συχνότητες σε τάξεις, πόσες φορές δηλ. εμφανίζεται ο

βαθμός 0, 1, 2, ..., 99, 100 στο συνολικό δείγμα. Έτσι η πρώτη γραμμή του Πίνακα 3.1 μας λέει ότι μεταξύ των 43 παρατηρήσεων - βαθμολογία υπαλλήλων, 5 πήραν βαθμό από 0 έως και 40.

Ο αριθμός των παρατηρήσεων ή συχνοτήτων f_t σε κάθε τιμή της μεταβλητής ή σε κάθε τάξη της, ονομάζεται απόλυτη συχνότητα της τιμής, ενώ η τιμή του πηλίκου f/Σf ονομάζεται σχετική συχνότητα.

Πίνακας 3.1

Βαθμολογία Υπαλλήλων του Υπουργείο «Χ» στο Τεστ Δεξιοτήτων		
Τάξεις	Κεντρική Τιμή	Συχνότητες f_i
[0, 40]	20	$f_1 = 5$
[41, 80]	60	$f_2 = 15$
[81, 100]	90	$f_3 = 23$
Σύνολο		$\sum_{i=1}^{3} f_i = 43$

3.3.2 Η περίπτωση της συνεχούς μεταβλητής

Ορίσαμε νωρίτερα τη συνεχή μεταβλητή ως αυτή που μπορεί να λάβει κάθε τιμή από το διάστημα τιμών στο οποίο ορίζεται. Έτσι σε αντιδιαστολή με την ασυνεχή μεταβλητή που μπορεί να παρουσιαστεί με ή χωρίς τάξεις για τις τιμές των παρατηρήσεών της, η συνεχής μεταβλητή παρουσιάζεται αποκλειστικά με την ομαδοποίηση των τιμών σε τάξεις. Ακολουθούμε λοιπόν την ίδια τεχνική που αναφέρθηκε και πριν, δηλ. της ταξινόμησης των παρατηρήσεων σε τάξεις αφού χωρίσουμε το εύρος του διαστήματος των τιμών της μεταβλητής σε ορισμένο αριθμό διαδοχικών διαστημάτων. Ακολούθως σε κάθε διάστημα-τάξη τοποθετούμε τον αριθμό των παρατηρήσεων-συχνοτήτων που αντιστοιχεί σ' αυτό (στο αντίστοιχο διάστημα).

Ένα θέμα που ανακύπτει με τις τάξεις των τιμών των παρατηρήσεών μας είναι αυτό που αφορά το εύρος τους. Οι τάξεις μπορούν να έχουν όλες το ίδιο ακριβώς εύρος ή να διαφέρουν μεταξύ τους. Η θεωρία της στατιστικής ανάλυσης δεν υποδεικνύει ένα και μόνο τρόπο ως τον καταλληλότερο αλλά μας δίνει τη ευχέρεια να κινηθούμε ανάλογα με το στόχο της ανάλυσης. Έτσι εάν έχουμε παρατηρήσεις των οποίων οι τιμές τείνουν να συγκεντρώνονται σε συγκεκριμένες περιοχές και να είναι αραιότερες σε άλλες, μπορούμε να έχουμε μικρότερου εύρους τάξεις στις περιοχές με μεγάλη συγκέντρωση και μεγαλύτερου εύρους σε εκείνες τις περιοχές με τη μικρότερη συγκέντρωση. Ένα παράδειγμα είναι το ύψος του δηλωθέντος φορολογητέου εισοδήματος για ένα οικονομικό έτος στην Ελλάδα. Η μεταβλητή αυτή όπως είναι φυσικό λαμβάνει τιμές από 0 έως το άπειρο. Δεν υπάρχει δηλαδή κλειστό πεδίο τιμών. Επειδή τα περισσότερα εισοδήματα που δηλώνονται είναι από 0€ έως 60.000€ θα μας ενδιέφερε να αναλύσουμε αυτό το τμήμα του πεδίου τιμών της μεταβλητής που αποτελεί και περί το 80 % των υποβληθεισών φορολογικών δηλώσεων. Έτσι δημιουργούμε τάξεις π.χ. από 0 έως €10.000, από €10.001 έως 20.000€ κ.λπ. Για τα μεγάλα εισοδήματα που είναι μόνο το 20% του συνολικού

πληθυσμού μας μπορούμε να δημιουργήσουμε μία ή δύο τάξεις π.χ. από 60.001€ έως 100.000€ και από 100.001€ έως το άπειρο.

Σε κάθε διάστημα βέβαια θα υπολογίσουμε, όπως αναφέρθηκε προηγουμένως, την κεντρική τιμή του διαστήματος με τη διαδικασία που αναφέραμε. Θα πρέπει να προσέξουμε όμως ότι η τελευταία τάξη στο παράδειγμά μας είναι ανοικτή από πάνω που σημαίνει ότι δεν μπορούμε να υπολογίσουμε κεντρική τιμή. Αυτό δεν αποτελεί ιδιαίτερο πρόβλημα στο παρόν στάδιο. Θα δούμε όμως αργότερα ότι μας περιορίζει στον υπολογισμό κάποιων παραμέτρων τάσεως.

Θα πρέπει στο σημείο αυτό να αναφέρουμε ότι δεδομένης της έλλειψης υπόδειξης από τη θεωρεία της στατιστικής ανάλυσης ενός και μόνου ή κάποιων τρόπων ως τους καταλληλότερους για την εύρεση του αριθμού των τάξεων που θα πρέπει να χρησιμοποιήσουμε και του εύρους που αυτές θα πρέπει να έχουν, στην πράξη συχνά εφαρμόζεται ο εμπειρικός κανόνας του Stugres. Με βάση αυτόν τον κανόνα αυτό το πλάτος των τάξεων και ο αριθμός (πλήθος) αυτών υπολογίζονται ως εξής:

1. Βρίσκουμε την αριθμητική απόκλιση μεταξύ της μικρότερης και της μεγαλύτερης από τις τιμές των παρατηρήσεων της μεταβλητής μας.

Εύρος (R) = Xmax-Xmin

2. Υπολογίζουμε τον αριθμό των τάξεων: Πλήθος Τάξεων =1+3,33λογ(n), όπου n είναι το μέγεθος του δείγματος (ή Ν εάν πρόκειται για τον πληθυσμό). Το αποτέλεσμα μπορεί να μην είναι ακέραιος αριθμός και σε αυτήν την περίπτωση στρογγυλοποιούμε στον πλησιέστερο ακέραιο.

3. Υπολογίζουμε το πλάτος των τάξεων διαιρώντας το Εύρος με το πλήθος των τάξεων. δ=Εύρος / Πλήθος Τάξεων.

3. 4 Η γραφική παρουσίαση της κατανομής

Μετά τη δημιουργία της κατανομής συχνότητας και για να είναι ολοκληρωμένη η εικόνα που έχουμε για τη μεταβλητή μας μπορούμε να την παρουσιάσουμε και με μια γραφική αναπαράσταση. Η στατιστική ανάλυση μας δίνει τη δυνατότητα να χρησιμοποιήσουμε τα γραφικά εργαλεία του ιστογράμματος, του πολύγωνου συχνοτήτων ή αλλιώς πολυγωνική γραμμής των απολύτων ή σχετικών συχνοτήτων, καθώς επίσης και τη διαγραμματική παρουσίαση της αθροιστικής κατανομής συχνοτήτων.

3.4.1 Το Ιστόγραμμα Συχνοτήτων

Εάν σε ένα σύστημα κάθετων αξόνων χρησιμοποιήσουμε ορθογώνια παραλληλόγραμμα με εύρος ίσο με το εύρος κάθε τάξης των τιμών των παρατηρήσεών μας και με ύψος ίσο με τη συχνότητα της κάθε τάξης έχουμε το ιστόγραμμα συχνοτήτων.

3.4.2 Η Πολυγωνική Γραμμή ή Πολύγωνο Συχνοτήτων

Στο παραπάνω ιστόγραμμα συχνοτήτων εάν ενώσουμε μεταξύ του τα σημεία που προσδιορίζονται ως το μέσον της επάνω βάσης κάθε ορθογώνιου παραλληλόγραμμου

σχηματίζουμε μία πολυγωνική γραμμή. Προκειμένου η γραμμή αυτή να εφάπτεται στα δύο άκρα της επάνω στον οριζόντιο άξονα θα πρέπει να υποθέσουμε την ύπαρξη δύο εικονικών τάξεων με μηδενική συχνότητα πριν την πρώτη και μετά την τελευταία τάξη. Η κατασκευή που έχουμε τώρα είναι η πολυγωνική γραμμή ή όπως αλλιώς λέγεται το πολύγωνο συχνοτήτων[2]:

Παρουσιάζουμε παρακάτω ένα παράδειγμα. Στον πίνακα 3.2 έστω ότι έχουμε την κατανομή των μηνιαίων αποδοχών 150 υπαλλήλων ενός υπουργείου.

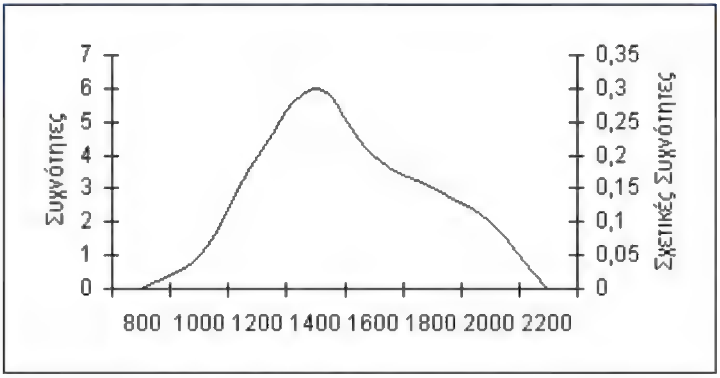
Πίνακας 3.2

Τάξεις εισοδήματος σε €	x_i	f_i	f_i %	F_i	F_i %
<900	-	0	0	0	0
[900, 1100)	1000	1	0,05	1	0,05
[1100, 1300)	1200	4	0,2	5	0,25
[1300, 1500)	1400	6	0,3	11	0,55
[1500, 1700)	1600	4	0,2	15	0,75
[1700, 1900)	1800	3	0,15	18	0,9
[1900, 2100)	2000	2	0,1	20	1
≥2100	-	0	0	20	1
Σύνολα		**20**	**1**		

Το ιστόγραμμα συχνοτήτων του Πίνακα 3.2 παρουσιάζεται στο Διάγραμμα 3.1.

ΔΙΑΓΡΑΜΜΑ 3.1

Μηνιαίες Αποδοχές Υπαλλήλων του Υπουργείου «Χ»



[2] Είναι προφανές ότι η μορφή του ιστογράμματος επηρεάζεται δραστικά από την επιλογή των κλάσεων.

Επισημαίνουμε ότι κάθε ορθογώνιο του ιστογράμματος σχεδιάζεται έτσι, ώστε, το εμβαδόν του να ισούται με τη συχνότητα (ή τη σχετική συχνότητα) της αντίστοιχης κλάσης[3]. Επομένως το συνολικό εμβαδόν των ορθογωνίων είναι ίσο με το πλήθος των παρατηρήσεων Ν (ή είναι ίσο με 1). Επίσης, το εμβαδόν που περικλείεται μεταξύ του πολυγώνου συχνοτήτων ή σχετικών συχνοτήτων και του οριζόντιου άξονα είναι ίσο με Ν ή με 1 αντίστοιχα. Οποιοδήποτε τμήμα αυτού του εμβαδού μπορεί να υπολογιστεί (ακριβέστερα, να εκτιμηθεί), δίνοντάς μας το ποσοστό των παρατηρήσεων που βρίσκονται μεταξύ δύο τιμών της μεταβλητής ή αριστερά μιας τιμής ή δεξιά μιας τιμής. Η παραπάνω τεθλασμένη γραμμή, αποτελεί τη λεγόμενη πολυγωνική γραμμή ή πολύγωνο συχνοτήτων. Αν το πλάτος των κλάσεων είναι πολύ μικρό το πολύγωνο συχνοτήτων παίρνει μορφή λείας καμπύλης η οποία ονομάζεται καμπύλη συχνοτήτων.

3.5 Η αθροιστική κατανομή συχνοτήτων: η έννοια των κατανομών «μικρότερη από» και «μεγαλύτερη από»

Πέραν αυτών που αναφέρθηκαν προηγούμενα υπάρχουν δύο ακόμη τύποι κατανομής συχνοτήτων. Η κατανομή «μικρότερη από» και η κατανομή «μεγαλύτερη από». Στον Πίνακα 3.3 παρουσιάζουμε αυτούς τους δύο τύπους.

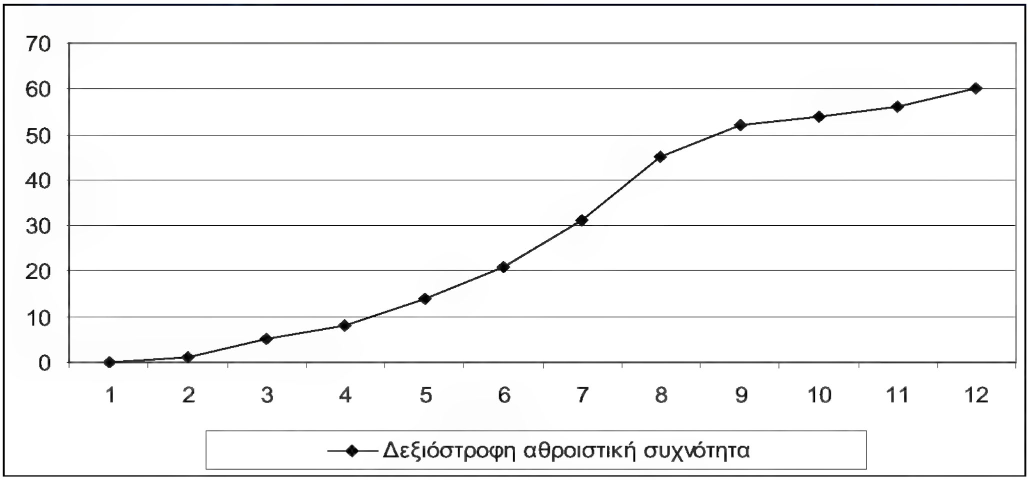
Πίνακας 3.3

Τιμή Ορισμένων Μετοχών την 31/12/08 στο Χρηματιστήριο Αθηνών			
		Αθροιστική Κατανομή	
Τιμή μετοχών	Αριθμός μετοχών	«Μικρότερη από» Δεξιόστροφη αθροιστική συχνότητα (F_i)	Μεγαλύτερη από» Αριστερόστροφη αθροιστική συχνότητα (Φ_i)
8,00	0	0	60
8,50	1	1	59
9,00	4	5	55
9,50	3	8	52
10,00	6	14	46
10,50	7	21	35
11,00	10	31	29
11,50	14	45	15
12,00	7	52	8
12,50	2	54	6
13,00	2	56	4
16,00	4	60	0

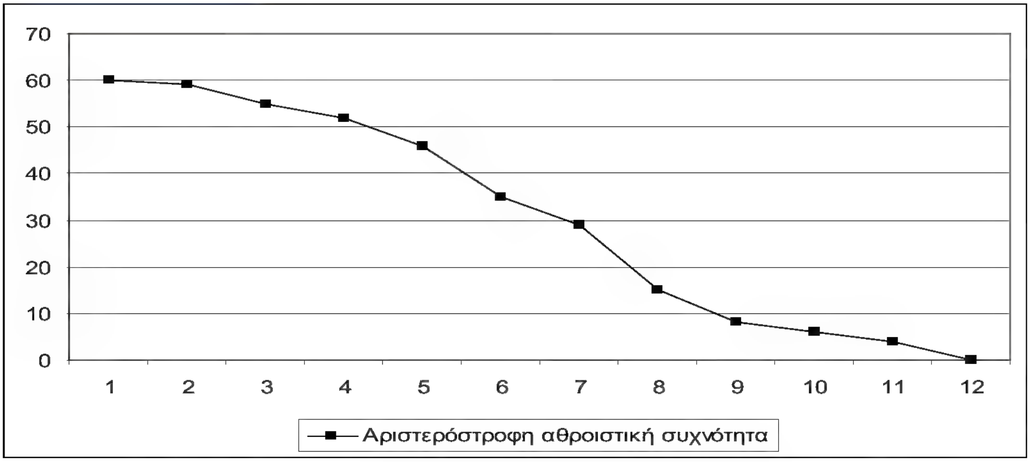
[3] Αν όλες οι κλάσεις έχουν ίδιο πλάτος, τότε προφανώς και τα ύψη των ορθογωνίων θα είναι ίσα με τις αντίστοιχες συχνότητες ή σχετικές συχνότητες. Αν όμως οι κλάσεις δεν έχουν ίδιο πλάτος τότε μόνο τα εμβαδά είναι ίσα με τις αντίστοιχες συχνότητες ή τις σχετικές συχνότητες και όχι τα ύψη.

Στην περίπτωση της «μικρότερη από» αθροιστικής κατανομής εμφανίζεται το άθροισμα των συχνοτήτων που αντιστοιχούν σε τιμές της μεταβλητής μικρότερες από μια συγκεκριμένη τιμή. Αντίθετα στην περίπτωση «μεγαλύτερη από» εμφανίζεται το άθροισμα των συχνοτήτων που αντιστοιχούν σε τιμές της μεταβλητής μεγαλύτερες από μια συγκεκριμένη τιμή. Μια παραλλαγή των δύο αυτών μορφών είναι οι κατανομές «μικρότερη ή ιση» ή «μεγαλύτερη ή ιση» που συγκρίνουν τιμές της μεταβλητής από μία συγκεκριμένη τιμή και πάνω ή όχι. Στο Διάγραμμα 3.2 και 3.3 παρουσιάζονται οι κατανομές του Πίνακα 3.3.

ΔΙΑΓΡΑΜΜΑ 3.2



ΔΙΑΓΡΑΜΜΑ 3.3



Η πληροφορία που λαμβάνουμε από αυτό το είδος των κατανομών είναι ότι μας ενημερώνουν πόσες από τις τιμές των επιλεγμένων μετοχών την 31/12/08, είναι πάνω ή κάτω κάποιας συγκεκριμένης τιμής. Π.χ. από τα στοιχεία του πίνακα προκύπτει ότι 14 από τις 60 μετοχές έκλεισαν με τιμή μικρότερη ή ίση από 10,00€, ή 35 μετοχές έκλεισαν με τιμή μεγαλύτερη ή ίση από 10,50 €, κ.λπ.

3.6 Η σχετική κατανομή συχνότητας και η σχετική αθροιστική κατανομή συχνότητας

Παρουσιάσαμε προηγούμενα την έννοια της κατανομής συχνότητας η οποία δείχνει τον αριθμό των παρατηρήσεων σε κάθε μια από τις τιμές ή τα διαστήματα τιμών της μεταβλητής. Το άθροισμα των συχνοτήτων σε όλες τις τιμές ή διαστήματα τιμών ισούται με το σύνολο των παρατηρήσεων στο δείγμα ή στον πληθυσμό. Εάν τώρα προχωρήσουμε ένα βήμα ακόμη και διαιρέσουμε τη συχνότητα κάθε τιμής ή κάθε διαστήματος τιμών της μεταβλητής μας με το σύνολο των παρατηρήσεων μας (σύνολο δείγματος ή πληθυσμού) λαμβάνουμε το μέγεθος που ονομάζεται σχετική συχνότητα. Η ίδια διαδικασία ακολουθείται και στην περίπτωση αθροιστικής κατανομής συχνοτήτων προκειμένου να λάβουμε τις σχετικές αθροιστικές συχνότητες. Στους παρακάτω πίνακες παρουσιάζεται ένα παράδειγμα των σχετικών συχνοτήτων για ένα δείγμα 380 φορολογουμένων.

Πίνακας 3.4

Κατανομή Συχνότητας και Σχετική Κατανομή Συχνότητας		
Δηλωθέν Φορολογητέο Εισόδημα για το Οικονομικό Έτος 2008		
Φορολογητέο Εισόδημα	Συχνότητα	Σχετική Συχνότητα
<10.000	46	0,12
[10.001, 20.000)	106	0,28
[20.001, 30.000)	95	0,25
[30.001, 40.000)	53	0,14
[40.001, 60.000)	42	0,11
[60.001, 200.000)	23	0,06
[200.001, 1.000.000)	15	0,04
Σύνολο	380	1,00

Σημ.: ποσά σε ευρώ

Πίνακας 3.5

Αθροιστική Κατανομή Συχνότητας (Απόλυτη και Σχετική)		
Δηλωθέν Φορολογητέο Εισόδημα για το Οικονομικό Έτος 2008		
Σύνορα Τάξεων	Απόλυτη Αθροιστική Συχνότητα (F_i)	Σχετική Αθροιστική Συχνότητα (F_i%)
< 10.000	46	0,12
[10.001, 20.000]	152	0,40
[20.001, 30.000]	247	0,65
[30.001, 40.000]	300	0,79
[40.001, 60.000]	342	0,90
[60.001, 200.000]	365	0,96
[200.001, 1.000.000]	380	1,00
Ποσά σε ευρώ		

3.7 Παραδείγματα

Παράδειγμα 1

Για τις παρακάτω παρατηρήσεις ζητείται να ταξινομηθούν σε στατιστικούς πίνακες σχετικών συχνοτήτων και δεξιόστροφων αθροιστικών σχετικών συχνοτήτων. Να εξηγήσετε σύντομα τη χρησιμοποιούμενη μεθοδολογία σας.

8	5	9	10	7
1	4	5	6	9
2	7	11	4	6
3	5	6	8	8

Απάντηση:

Για την ταξινόμηση των παρατηρήσεων χρησιμοποιείται ο εμπειρικός Κανόνας του Stugres: δ=Εύρος / Πλήθος Τάξεων

Εύρος (R)=Xmax-Xmin=11-1=10

Πλήθος Τάξεων =1+3,33λογ(n)=5,33 ≈ 5. Επομένως δ=10/5=2

Τάξεις Xi	Απόλ.Συχν.(fi)	Σχετ.Συχν.(fi%)	Δεξ. Αθρ. Σειρά	Δεξ. Αθρ. Σχετ. Συχν. (Fi%)
[1 , 3]	3	0.15	μέχρι 3	0.15
(3 , 5]	6	0.3	μέχρι 5	0.45
(5 , 7]	5	0.25	μέχρι 7	0.7
(7 , 9]	4	0.2	μέχρι 9	0.9
(9 . 11]	2	0.1	μέχρι 11	1
Σύνολο	20	1		

Παράδειγμα 2

Η εταιρία ENRON A.E. χορήγησε σε 5000 στελέχη της διάφορα δάνεια για αγορά κατοικίας. Σε μια μελέτη που έγινε, ελήφθη δείγμα τριακοσίων δανείων το οποίο παρουσιάζεται στον παρακάτω πίνακα κατανομής. Να κατασκευάσετε ένα πίνακα που να εμφανίζει την απόλυτη συχνότητα και τη δεξιόστροφη απόλυτη συχνότητα.

Κατανομή Δανείων	
Όρια Δανείων	**Αριθμός Πελατών**
< 10.000	49
[10.000, 20.000)	90
[20.000, 30.000)	78
[30.000, 40.000)	42
[40.000, 50.000)	26
[50.000, 60.000)	15

Σημ.: Ποσά σε Ευρώ

Απάντηση:

Τάξεις	Απόλυτη Συχνότητα f_I	Δεξιόστροφη Αθροιστική Σειρά	Δεξιόστροφη Αθρ/κή Απόλυτη Συχνότητα F_I
(0, 10000]	49	έως 10	49
(10000, 20000]	90	έως 20	139
(20000, 30000]	78	έως 30	217
(30000, 40000]	42	έως 40	259
(40000, 50000]	26	έως 50	285
(50000, 60000]	15	έως 60	300
Σύνολο	*300*		

Παράδειγμα 3

Τριάντα δύο ελεγκτές χρειάζονται τους παρακάτω χρόνους σε ημέρες προκειμένου να ελέγξουν τα παραστατικά της εταιρίας του Θέματος 1 που πτώχευσε λόγω του υπερβολικού αριθμού δανείων που είχε χορηγήσει: 21, 20, 10, 16, 15, 9, 12, 11, 14, 12, 18, 15, 10, 16, 15, 14, 9, 11, 17, 12, 13, 11, 14, 13, 16, 14, 17, 13, 14, 11, 12, 15. Για τις παρατηρήσεις αυτές ζητείται να ταξινομηθούν σε τάξεις ίσου πλάτους (κατανομή απόλυτων και σχετικών συχνοτήτων) και να υπολογιστεί η αθροιστική συχνότητα.

Απάντηση:

Για την ταξινόμηση των παρατηρήσεων χρησιμοποιείται ο εμπειρικός Κανόνας του Stugres: δ=Εύρος/Πλήθος Τάξεων. Εφαρμόζοντας τον κανόνα έχουμε:

δ=(21-9)/1+3.33log32=12/6,01=2 δηλαδή θα κατασκευάσουμε έξι τάξεις με πλάτος 2.

X	f_i	f_i%	F_i
[9, 11)	4	0,125	4
[11, 13)	8	0,25	12
[13, 15)	8	0,25	20
[15, 17)	7	0,21875	27
[17, 19)	3	0,09375	30
[19, 21]	2	0,0625	32
Σύνολο	32	1,00	

4. Εκπαιδευτική Ενότητα

- **Περιγραφική Στατιστική ΙΙ: μέτρα συμπύκνωσης των πληροφοριών**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να υπολογίζει και να ερμηνεύει τα στατιστικά μέτρα τάσης ή θέσης (αταξινόμητων και ταξινομημένων δεδομένων): μέσοι όροι, διάμεσος, τεταρτημόρια, επικρατούσα τιμή.
- Να υπολογίζει και να ερμηνεύει τα στατιστικά μέτρα διασποράς: εύρος, διακύμανση, τυπική απόκλιση, συντελεστής μεταβλητότητας.
- Να υπολογίζει και να ερμηνεύει τα στατιστικά μέτρα ασυμμετρίας και κύρτωσης: συντελεστές ασυμμετρίας και κύρτωσης.
- Να αξιολογεί τη δομή μονομεταβλητών πληθυσμών με περιγραφικά στατιστικά εργαλεία ανάλυσης.
- Να εφαρμόζει τα παραπάνω χρησιμοποιώντας το στατιστικό πακέτο SPSS.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Αριθμητικός Μέσος
- Σταθμικός μέσος αριθμητικός
- Γεωμετρικός Μέσος
- Διάμεσος
- Τεταρτημόρια
- Επικρατούσα Τιμή
- Εύρος Διασποράς
- Μέση Απόκλιση
- Διακύμανση/Τυπική Απόκλιση
- Συντελεστής Μεταβλητότητας
- Παράμεροι Ασυμμετρίας
- Παράμετροι Κύρτωσης

Σε προηγούμενο κεφάλαιο αναφέραμε ότι η στατιστική ανάλυση αποτελεί μια διαδικασία που περιλαμβάνει πέντε συγκεκριμένα στάδια:

Στάδιο πρώτο: Η συλλογή των στατιστικών στοιχείων.

Στάδιο δεύτερο: Η οργάνωση των στατιστικών στοιχείων.

Στάδιο τρίτο: Η παρουσίαση των στατιστικών στοιχείων.

Στάδιο τέταρτο: Η ανάλυση των στατιστικών στοιχείων και εξαγωγή στατιστικών συμπερασμάτων.

Στάδιο πέμπτο: Η εφαρμογή των συμπερασμάτων του τέταρτου σταδίου στο σύνολο των μονάδων του πληθυσμού.

Μέχρι τώρα έχουμε παρουσιάσει τον τρόπο που δουλεύουμε στα πρώτα τρία στάδια. Στο κεφάλαιο αυτό θα παρουσιάσουμε τα εργαλεία και τη μεθοδολογία προκειμένου να λειτουργήσουμε στο τέταρτο στάδιο. Αυτά είναι οι λεγόμενες Παράμετροι (ή Στατιστικές στην περίπτωση των δειγμάτων) Τάσης, Θέσης, Διασποράς, Ασυμμετρίας και Κύρτωσης.

Σε αντιδιαστολή με ότι κάναμε στα προηγούμενα στάδια όπου συνοψίζαμε το πλήθος των πρωτογενών αριθμητικών στοιχείων σε απλούς πίνακες και διαγράμματα, εδώ θα συμπτύξουμε παραπέρα τις πληροφορίες των δεδομένων μας, προκειμένου να έχουμε με τη βοήθεια των παραμέτρων μια όσο το δυνατόν πιο ικανοποιητική και ακριβή περιγραφή των κύριων χαρακτηριστικών του υπό διερεύνηση πληθυσμού ή δείγματος.

4.1 Οι παράμετροι τάσης

Μια πρώτη προσέγγιση που μας είναι απαραίτητη κατά τη μελέτη των στατιστικών μας στοιχείων είναι να δούμε κατά πόσον αυτά συγκεντρώνονται σε ένα συγκεκριμένο σημείο. Στην προσέγγιση αυτή μας βοηθούν οι λεγόμενες παράμετροι τάσης.

Οι παράμετροι αυτοί είναι ο μέσος αριθμητικός, ο γεωμετρικός μέσος και ο αρμονικός μέσος. Ο τελευταίος δεν χρησιμοποιείται σχεδόν ποτέ στη στατιστική ανάλυση.

4.1.1 Ο Αριθμητικός Μέσος

Το επίπεδο των υπαλλήλων ενός υπουργείου εκτιμάται με βάση τη βαθμολογία που πετυχαίνουν στο τεστ δεξιοτήτων κάθε έτος. Προκειμένου να έχουμε μια εικόνα που περίπου βρίσκεται κατά προσέγγιση το επίπεδο των υπαλλήλων χρειαζόμαστε τη μέση βαθμολογία που αυτοί λαμβάνουν στο τεστ δεξιοτήτων. Ο μέσος αυτός όρος του πληθυσμού (υπάλληλοι του υπουργείου) που μας φανερώνει ένα χαρακτηριστικό (επίπεδο υπαλλήλων) καλείται και μέσος πληθυσμού. Για τον υπολογισμό του υπάρχουν δύο τύποι. Ο πρώτος αφορά την περίπτωση του αστάθμητου μέσου αριθμητικού και εφαρμόζεται στην περίπτωση που τα δεδομένα μας δεν είναι ταξινομημένα. Ο δεύτερος αφορά την περίπτωση του σταθμικού αριθμητικού μέσου και εφαρμόζεται όταν τα δεδομένα μας είναι ταξινομημένα σε τάξεις. Έτσι έχουμε:

- για την περίπτωση των αταξινόμητων δεδομένων,

Αστάθμητος μέσος αριθμητικός: $\bar{x} = \dfrac{\sum_{i=1}^{N} x_1}{N}$	**(4.1)**

όπου x_1, x_2,..., x_N οι τιμές της μεταβλητής X και N το πλήθος των παρατηρήσεων του πληθυσμού.

- για την περίπτωση κατά την οποία τα στατιστικά στοιχεία δίνονται σε μορφή κατανομής συχνοτήτων, για τον υπολογισμό του αριθμητικού μέσου χρησιμοποιούμε ως συντελεστές στάθμισης τις συχνότητες της μεταβλητής. Ο τύπος του σταθμικού αριθμητικού μέσου, όπως ονομάζεται στην περίπτωση αυτή ο αριθμητικός μέσος, έχει ως εξής:

Σταθμικός μέσος αριθμητικός (άμεσος υπολογισμός):

$\bar{x} = \dfrac{f_1x_1 + f_2x_2 + \ldots + f_kx_k}{f_1 + f_2 + \ldots + f_k} = \dfrac{\sum f_i x_i}{\sum f_i}$	**(4.2)**

όπου x_1, x_2, x_3..., x_k οι τιμές της μεταβλητής X και f_1, f_2,..., f_k οι συχνότητες των τιμών της μεταβλητής X.

Στην περίπτωση δηλαδή που για τη μεταβλητή μας υπάρχουν τάξεις ή διαστήματα τιμών της μεταβλητής, στις τιμές x_1, x_2, x_3..., x_k του τύπου (4.2) χρησιμοποιούμε την κεντρική τιμή της τάξης ή του διαστήματος τιμών της μεταβλητής (έμμεσος υπολογισμός).

Τέλος θα πρέπει να αναφερθεί ότι υπάρχει και αριθμητικός μέσος των αριθμητικών μέσων k δειγμάτων μεγέθους n_1, n_2,..., n_k, αντίστοιχα, άκυρο είναι:

$\bar{\bar{x}} = \dfrac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$	**(4.3)**

που είναι επί της ουσίας ένας σταθμικός μέσος αριθμητικός.

Σημειώνουμε τέλος ότι ο μέσος αριθμητικός που αφορά τον πληθυσμό συμβολίζεται με μ.

Παράδειγμα 4.1

Ένας οδηγός φορτηγού διανομής τροφίμων, αγόρασε σε μια ημέρα πετρέλαιο από τρία διαφορετικά πρατήρια. Από το πρώτο αγόρασε 6 λίτρα προς 0,75€ το λίτρο, από το δεύτερο 12 λίτρα προς 0,84€ το λίτρο και από το τρίτο 5 λίτρα προς 0,76€ το λίτρο. Ποια είναι η μέση τιμή που πλήρωσε;

Απάντηση:

Για να υπολογιστεί το μέσο ποσό που πλήρωσε ανά λίτρο ο οδηγός πρέπει να χρησιμοποιηθεί ο σταθμικός μέσος:

$$\overline{\chi}_w = \frac{\sum_{i=1}^{\nu} w_i \chi_i}{\sum_{i=1}^{\nu} w_i} = \frac{6 \cdot 0{,}75 + 12 \cdot 0{,}84 + 5 \cdot 0.76}{6+12+5} = 0.799\,€ \quad \text{ανά λίτρο}$$

Παράδειγμα 4.2

Αν το μέσο ύψος 10 φοιτητών είναι 170 cm και το μέσο ύψος 5 φοιτητριών είναι 160 cm, ποιο είναι το μέσο ύψος φοιτητών και φοιτητριών;

Απάντηση:

Το μέσο ύψος φοιτητών και φοιτητριών είναι:

$$\overline{\chi} = \frac{\sum_{i=1}^{2} n_i \overline{\chi}_i}{\sum_{i=1}^{2} n_i} = \frac{10 \cdot 170 + 5 \cdot 160}{15} = 166{,}7\,cm$$

4.1.2 Γεωμετρικός Μέσος

Ο γεωμετρικός μέσος μιας σειράς παρατηρήσεων X_1, X_2, X_3, ..., X_N ορίζεται ως η νιοστή ρίζα του γινομένου τους

$G = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$	**(4.4)**

Ο υπολογισμός του τύπου (4) γίνεται με λογαρίθμιση όπως φαίνεται παρακάτω:

$\lambda o\gamma G = \frac{\lambda o\gamma x_1 + \lambda o\gamma x_2 + \ldots + \lambda o\gamma x_N}{N} = \frac{\sum_{i=1}^{N} \lambda o\gamma x_i}{N}$	**(4.5)**

Σε περίπτωση συνεχούς μεταβλητής, λαμβάνονται υπ' όψιν οι συχνότητες ως σταθμίσεις και ο τύπος (4.5) γίνεται:

$G = \sqrt[\sum f]{x_1^{f} \cdot x_2^{f} \cdot \ldots \cdot x_N^{f}}$ ή $\lambda o \gamma G = \dfrac{\sum_{i=1}^{N} f_i \lambda o \gamma x_i}{\sum f_i}$	**(4.6)**

Η ερμηνεία του μέσου γεωμετρικού είναι ανάλογη του αριθμητικού. Έτσι, αν αντικατασταθούν οι τιμές της μεταβλητής από το μέσο γεωμετρικό, το γινόμενο τους θα παραμείνει αμετάβλητο.

4.2 Οι παράμετροι θέσης

Συνεχίζοντας την ανάλυση των στατιστικών μας στοιχείων και αφού υπολογίσαμε τις παραμέτρους τάσης, ενδιαφερόμαστε να ορίσουμε τη θέση της κατανομής. Σε αυτή τη διαδικασία χρησιμοποιούμε τις παραμέτρους της Διάμεσου, των Τεταρτημορίων, Δεκατημορίων, Εκατοστημορίων και της Επικρατούσας Τιμής. Η τελευταία παράμετρος συναντάται με τις ονομασίες Σημείο Μέγιστης Συχνότητας.

4.2.1 Η Διάμεσος

Όπως εξηγεί και η ετυμολογία της λέξης η διάμεσος είναι μια τιμή της μεταβλητής μας πάνω και κάτω από την οποία βρίσκεται ο ίδιος αριθμητικά αριθμός παρατηρήσεων του δείγματος. Προϋπόθεση για να ισχύει ο παραπάνω ορισμός είναι οι παρατηρήσεις μας να έχουν τοποθετηθεί σε αύξουσα σειρά. Ο υπολογισμός της διαμέσου ακολουθεί την παρακάτω διαδικασία:

- Σε περίπτωση περιττού αριθμού παρατηρήσεων, η διάμεσος ορίζεται επακριβώς ως η παρατήρηση της θέσης:

$\dfrac{N+1}{2}$	**(4.7)**

- Σε περίπτωση αρτίου αριθμού παρατηρήσεων ως ο μέσος των δύο κεντρικών διαδοχικών τιμών.

- Σε στοιχεία συνεχούς κατανομής η διάμεσος δίνεται από τον τύπο:

$M_e = x_i + \dfrac{\delta}{f_i}\left(\dfrac{N}{2} - F_i\right)$	**(4.8)**

όπου M_e διάμεσος, x_i το κάτω όριο του διαστήματος τιμών που περιέχεται η M_e, N ο

αριθμός των παρατηρήσεων στον πληθυσμό, F_i η αμέσως προηγούμενη του $N/2$ δεξιόστροφη αθροιστική συχνότητα, δ το πλάτος του διαστήματος τιμών (τάξεως) που περιέχεται η M_e, και f_i η απόλυτη συχνότητα του διαστήματος τιμών που περιέχεται η M_e.

Παράδειγμα 4.3

Το υπόλοιπο καταναλωτικών δανείων της Τράπεζας Α την 31-12-08 παρουσιάζεται παρακάτω ως εξής:

Ποσό Δανείου (€)	Συχνότητα Δανείων	Δεξιόστροφη Αθροιστική
[0, 50.000)	180	180
[50.001, 100.000)	110	290
[100.001, 250.000)	100	390
[250.001, 500.000)	90	480
[500.001, 1.000.000)	70	550
[1.000.001, 4.000.000)	50	600
ΣΥΝΟΛΟ	600	

Να υπολογιστεί η διάμεσος τιμή των καταναλωτικών δανείων.

Απάντηση:

Με χρησιμοποίηση του παραπάνω τύπου:

α) προσδιορίζουμε το διάστημα στο οποίο βρίσκεται η διάμεσος: N/2 =600/2=300. Άρα η διάμεσος βρίσκεται στο τρίτο διάστημα τιμών [100.001, 250.000].

β) στη συνέχεια υπολογίζουμε τα μέτρα που συνθέτουν τον τύπο της διαμέσου:

x_i =100.001 , f_i =290 , f_i =100 , δ=150.000

γ) τέλος, εφαρμόζουμε στον τύπο της διαμέσου:

$$M_e = x_i + \frac{\delta}{f_i}(\frac{N}{2} - F_i) = 100.001 + \frac{150.000}{100}(300 - 290) = 115.001€$$

Επομένως, πάνω και κάτω από το ύψος των 115.001€ χορηγήθηκαν από 300 δάνεια αντίστοιχα.

4.2.2 Τα Τεταρτημόρια, Δεκατημόρια, Εκατοστημόρια

Με παραλλαγή του τύπου (4.8) της διαμέσου μπορούμε να εξαγάγουμε σημαντικές πληροφορίες από τα στατιστικά μας δεδομένα όπως:

- πάνω από ποια τιμή βρίσκεται το 75% των παρατηρήσεων του πληθυσμού (πρώτο τεταρτημόριο)
- πάνω από ποια τιμή βρίσκεται το 25% (τρίτο τεταρτημόριο)
- κάτω από ποια τιμή βρίσκεται το 25% (πρώτο τεταρτημόριο)
- κάτω από ποια τιμή βρίσκεται το 75% (τρίτο τεταρτημόριο).

Οι τιμές αυτές είναι δηλαδή οι θέσεις του πρώτου και του τρίτου τετάρτου των παρατηρήσεων του πληθυσμού.

Κατά αντιστοιχία μπορούμε να ορίσουμε τα δεκατημόρια και τα εκατοστημόρια.
Οι τύποι που εφαρμόζουμε παρουσιάζονται παρακάτω.

- Το πρώτο τεταρτημόριο δίνεται από τον τύπο:

$Q_1 = x_i + \frac{\delta}{f_i}(\frac{N}{4} - F_i)$	**(4.9)**

- Το Τρίτο Τεταρτημόριο Q_3 δίνεται από τον τύπο:

$Q_3 = x_i + \frac{\delta}{f_i}(\frac{3N}{4} - F_i)$	**(4.10)**

- Τα Δεκατημόρια εξάλλου, υπολογίζονται από τον τύπο:

$D_k = x_i + \frac{\delta}{f_i}(\frac{k \cdot N}{10} - F_i)$	**(4.11)**

όπου k=1,2,3,...,9

- Τα Εκατοστημόρια δίνονται από τον τύπο:

$C_k = x_i + \frac{\delta}{f_i}(\frac{k \cdot N}{100} - F_i)$	**(4.12)**

Παράδειγμα 4.4

Χρησιμοποιώντας τα στοιχεία του Παραδείγματος 4.3 να υπολογιστεί το πρώτο και τρίτο τεταρτημόριο, το τέταρτο δεκατημόριο και το ενενηκοστό δεύτερο εκατοστημόριο.

Απάντηση:

Το πρώτο τεταρτημόριο

$$Q_1 = x_i + \frac{\delta}{f_i}(\frac{N}{4} - F_i) = 0 + \frac{50000}{180}(180 - 0) = 41.667€$$

Δηλαδή πάνω από την τιμή Q_1=41.667€, χορηγήθηκε το 75% των δανείων και το υπόλοιπο 25% κάτω από αυτή.

Το τρίτο τεταρτημόριο

$$Q_3 = x_i + \frac{\delta}{f_i}(\frac{3 \cdot N}{4} - F_i) = 250.001 + \frac{250.000}{90}(450 - 390) = 416.668€$$

Δηλαδή πάνω από την τιμή αυτή χορηγήθηκαν 150 δάνεια ή το 25% των δανείων. Υπολογίζοντας το τέταρτο δεκατημόριο έχουμε την πληροφορία πάνω από ποια τιμή χορηγήθηκε το 60% των δανείων:

$$D_4 = 50.001 + \frac{50.000}{110}(240 - 180) = 77.274€$$

Η τιμή της μεταβλητής που υπολογίσαμε δηλώνει ότι κάτω από το ύψος των 77.274€ χορηγήθηκε το 40% των δανείων (240 δάνεια) και πάνω από αυτή την τιμή το 60%.

Τέλος, υπολογίζοντας το ενενηκοστό δεύτερο εκατοστημόριο βρίσκουμε ποια είναι η τιμή της μεταβλητής κάτω από την οποία περιλαμβάνεται το 92% των παρατηρήσεων

$$C_{92} = 1.000.001 + \frac{3.000.000}{50_i}(552 - 550) = 1.120.001€$$

Με άλλα λόγια, από τα 600 δάνεια που χορήγησε η Τράπεζα αυτή, τα 552 αφορούσαν ποσά μικρότερα του 1.120.001€.

Παράδειγμα 4.5

Στον παρακάτω πίνακα συχνοτήτων δίνεται η κατανομή της βαθμολογίας 50 φοιτητών της Εθνικής Σχολής Δημόσιας Διοίκησης. Αν στο 5% των φοιτητών με την υψηλότερη βαθμολογία δίνεται η δυνατότητα να διοριστούν στο υπουργείο της πρώτης τους επιλογής, τι βαθμό θα πρέπει να έχει ένας τέτοιος φοιτητής;

Βαθμός	x_i	f_i	F_i
[10, 12)	11	5	5
[12, 14)	13	10	15
[14, 16)	15	20	35
[16, 18)	17	10	45
[18, 20)	19	5	50

Απάντηση:

Αφού ψάχνουμε το 5% θα χρησιμοποιήσουμε τον τύπο του εκατοστημορίου.

$$C_{95} = x_i + \frac{\delta}{f_i}\left(\frac{k \cdot N}{100} - F_i\right) = 18 + \frac{2}{5}(47,5 - 45) \approx 19$$

Άρα η βαθμολογία που θα πρέπει να πετύχουν οι φοιτητές είναι τουλάχιστον 19 (κλίμακα 0-20).

4.2.3 Η Επικρατούσα Τιμή

Εάν μία τιμή της μεταβλητής μας εμφανίζει συχνότητα μεγαλύτερη από τις άλλες τότε αυτή η τιμή ορίζεται ως η τιμή με τη μέγιστη συχνότητα ή αλλιώς η επικρατούσα τιμή και συμβολίζεται με το M_0. Ο τύπος της είναι:

$$M_0 = x_i + \delta \cdot \frac{\Delta_1}{\Delta_1 + \Delta_2}$$	**(4.13)**

όπου, X_i = κατώτερο όριο τάξης στην οποία αντιστοιχεί η μεγαλύτερη συχνότητα (f_{max}), δ= το πλάτος της τάξης με τη μεγαλύτερη συχνότητα, $\Delta_1 = f_{max} - f_{max-1}$ και $\Delta_2 = f_{max} - f_{max+1}$

Σημειώνουμε εδώ ότι εάν δεν βρούμε μία και μόνη τιμή που να εμφανίζει τη μέγιστη συχνότητα αλλά υπάρχουν τουλάχιστον δύο τιμές με την ίδια συχνότητα (μεγαλύτερη όμως από όλες τις άλλες), τότε δεν υπάρχει επικρατούσα τιμή.

Παράδειγμα 4.6

Εφαρμόζοντας τον τύπο (4.13) για τα στοιχεία του Παραδείγματος 4.3 να υπολογίσετε την επικρατούσα τιμή.

Απάντηση:

$$M_0 = 0 + 50.000 \cdot \frac{(180-0)}{(180-0)+(180-110)} = 36.000€$$

Με άλλα λόγια το συνηθέστερα χορηγούμενο ύψος καταναλωτικού δανείου της Τράπεζας είναι M_0=36.000€.

Σημειώνουμε εδώ ότι στην περίπτωση που μια κατανομή είναι πλήρως συμμετρική, η επικρατούσα τιμή, η διάμεσος και ο μέσος αριθμητικός θα έχουν το ίδιο αριθμητικό αποτέλεσμα αφού συμπίπτουν στην ίδια θέση.

4.3 Οι παράμετροι διασποράς

Παρουσιάστηκαν παραπάνω παράμετροι τάσης (γύρω από ποια κεντρική τιμή τείνουν να συγκεντρώνονται τα στοιχεία μας;) και παράμετροι θέσης (πού βρίσκονται συγκεκριμένες τιμές της μεταβλητής μας;). Μας ενδιαφέρει ακόμη στη στατιστική ανάλυση να προσδιορίσουμε τι αποκλίσεις (διασπορά) εμφανίζουν τα στοιχεία μας από τις παραμέτρους τάσεις που υπολογίσαμε. Οι παράμετροι διασποράς που συνήθως χρησιμοποιούμε είναι το εύρος διασποράς, η μέση απόκλιση, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητικότητας. Η πληροφορία που λαμβάνουμε είναι χρήσιμη διότι μπορούμε να εκτιμήσουμε την αξιοπιστία των μέσων όρων ως αντιπροσωπευτικών τιμών της κεντρικής τάσης της μεταβλητής. Η ύπαρξη μεγάλης διασποράς περί το μέσο αριθμητικό προφανώς δεν τον καθιστά αξιόπιστη στατιστική πληροφορία.

4.3.1 Το Εύρος Διασποράς

Ως εύρος διασποράς ορίζουμε τη διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής των παρατηρήσεων του πληθυσμού ή του δείγματός μας. Έτσι έχουμε:

$R = x_{max} - x_{min}$	**(4.14)**

Ο υπολογισμός του εύρους διασποράς δεν μας δίνει σημαντικές πληροφορίες όταν έχουμε την περίπτωση πολύ απομακρυσμένων τιμών. Εάν τα στατιστικά μας δεδομένα είναι ταξινομημένα σε κατανομές συχνοτήτων, το εύρος ισούται με τη διαφορά του πάνω ορίου του τελευταίου διαστήματος τάξης, μείον το κάτω όριο του πρώτου διαστήματος τάξης. Είναι προφανές ότι εάν η τελευταία τάξη είναι ανοικτή στο άνω άκρο της δεν μπορώ να υπολογίσω το εύρος.

4.3.2 Η Μέση Απόκλιση

Η μέση απόκλιση ορίζεται ως ο μέσος αριθμητικός όλων των απολύτων διαφορών των τιμών μιας μεταβλητής από το μέσο αριθμητικό (μ) της μεταβλητής.

Ο τύπος δίδεται με δύο παραλλαγές για αταξινόμητα και ταξινομημένα δεδομένα:

Αταξινόμητα δεδομένα:

$M.A.=\dfrac{\sum\left\lvert x_i-\bar{x}\right\rvert}{N}$	**(4.15)**

Ταξινομημένα δεδομένα:

$M.A=\dfrac{\sum f_i\left\lvert x_i-\bar{x}\right\rvert}{\sum f_i}$	**(4.16)**

4.3.3 Η Διακύμανση και η Τυπική Απόκλιση

Η έννοια της διακύμανσης ορίζεται ως ο μέσος αριθμητικός των τετραγώνων των αποκλίσεων των τιμών της μεταβλητής από το μέσο αριθμητικό της, (μ), και συμβολίζεται με το γράμμα σ^2 (αναφερόμενη στην πληθυσμιακή διακύμανση). Ο τύπος δίδεται με δύο παραλλαγές:

- περίπτωση αταξινόμητων δεδομένων:

$\sigma^2=\dfrac{\sum(x_i-\bar{x})^2}{N}$	**(4.17)**

- περίπτωση ταξινομημένων στοιχείων σε μορφή κατανομής συχνοτήτων:

$\sigma^2=\dfrac{\sum f_i(x_i-\bar{x})^2}{N}$	**(4.18)**

Η τυπική απόκλιση ή μέση απόκλιση τετραγώνου ορίζεται ως η θετική τετραγωνική ρίζα της διακύμανσης και παρουσιάζεται με το ελληνικό γράμμα σ.

Ο τύπος δίνεται με δύο παραλλαγές:

- περίπτωση αταξινόμητων δεδομένων:

$\sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{N}}$	**(4.19)**

- περίπτωση ταξινομημένων στοιχείων σε μορφή κατανομής συχνοτήτων:

$\sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum f_i \cdot (x_i - \bar{x})^2}{N}}$	**(4.20)**

Στην περίπτωση που υπολογίζουμε τη διακύμανση ή την τυπική απόκλιση δειγματικών δεδομένων στους παραπάνω τύπους όπου: σ^2, σ, μ και Ν θα βάζουμε αντίστοιχα τα λατινικά γράμματα: s^2, s, $\bar{x}$,και n.

Στην περίπτωση που χρησιμοποιώντας στοιχεία δείγματος επιθυμούμε να εκτιμήσουμε (με κάποια πιθανότητα, όπως θα δούμε σε επόμενα κεφάλαια) τις αντίστοιχες πληθυσμιακές παραμέτρους (σ^2 ή σ) τότε οι στατιστικές του δείγματος (s^2 ή s) για να είναι οι αμερόληπτες (θα οριστεί σε μεταγενέστερο κεφάλαιο η έννοια αυτή) εκτιμήσεις τους, θα πρέπει να υπολογίζονται από τους αντίστοιχους τύπους, αντικαθιστώντας τον παρονομαστή με n-1 αντί του Ν.

Οι τύποι έχουν στον παρονομαστή όχι το πλήθος των παρατηρήσεων αλλά τους βαθμούς ελευθερίας τους, δηλαδή το πλήθος των πραγματικά ανεξάρτητων (χωρίς περιορισμούς) παρατηρήσεων.

Θα πρέπει να σημειώσουμε στο σημείο αυτό ότι κατά τον υπολογισμό της τιμής της διακύμανσης ή της τυπικής απόκλισης, ταξινομημένων δεδομένων σε κατανομές συχνότητας κάνουμε ένα στατιστικό σφάλμα. Το σφάλμα αυτό προέρχεται από τη διπλή αυθαίρετη υπόθεση σύμφωνα με την οποία αφενός θεωρούμε τους κεντρικούς όρους ως μέσους αριθμητικούς κάθε διαστήματος τάξεως, κι αφετέρου ότι οι συχνότητες (f_1) κάθε διαστήματος τάξεως είναι ομοιόμορφα κατανεμημένες γύρω από τον κεντρικό τους όρο ($\bar{x}$).
Έτσι η θεωρητική τιμή της διακύμανσης (από ταξινομημένα δεδομένα) θα διαφέρει οπωσδήποτε της πραγματικής (από τα αταξινόμητα πρωτογενή στοιχεία). Διόρθωση αυτού του σφάλματος έτσι ώστε η θεωρητική τιμή της διακύμανσης να προσεγγίζει την πραγματική της, δίνεται από το τύπο του W. F. Sheppard:

$\sigma_c^2 = \sigma^2 - \dfrac{\delta^2}{12}$	**(4.21)**

Η μέθοδος του Sheppard δεν συνιστάται σε εκείνες τις περιπτώσεις που έχουμε πολύ έντονα ασυμμετρικές κατανομές όπως π.χ. σχήματος U ή J ή όταν το μέγεθος του δείγματος είναι μεγαλύτερο από 500 παρατηρήσεις.

4.3.4. Ο Συντελεστής Μεταβλητικότητας

Κατά τη στατιστική επεξεργασία των στοιχείων μας δεν έχουμε πάντα μεταβλητές που εκφράζονται στις ίδιες μονάδες μέτρησης. Στις περιπτώσεις εκείνες που προσπαθούμε να συγκρίνουμε μεταβλητές ή κατανομές μεταβλητών οι οποίες εκφράζονται σε διαφορετικές μονάδες μέτρησης (χιλιόμετρα, ευρώ, εκατομμύρια κ.λπ.) η τυπική απόκλιση σαν μέτρο διασποράς παρουσιάζει αδυναμίες.

Στην περίπτωση αυτή μπορούμε να έχουμε μια αναγωγή των μεγεθών που μας ενδιαφέρουν σε ποσοστά ώστε μελετήσουμε τη σχετική και όχι την απόλυτη διασπορά που εκφράζει η τυπική απόκλιση. Το μέτρο αυτό της σχετικής διασποράς είναι ο συντελεστής μεταβλητικότητας. Η χρήση του μας εξασφαλίζει ανεξαρτησία από τις μονάδες μέτρησης που χρησιμοποιούμε. Εκφράζεται ως ο λόγος της τυπικής απόκλισης μιας κατανομής προς τον αριθμητικό μέσο της:

$$CV = \frac{\sigma}{\bar{x}} \cdot 100 \qquad \textbf{(4.22)}$$

Η ερμηνεία του συντελεστή είναι ότι όσο μικρότερη είναι η τιμή του τόσο πιο ομοιογενείς είναι οι τιμές της μεταβλητής.

Η χρήση του συντελεστή, π.χ., στην ανάλυση των τιμών των αγαθών από το Υπουργείο Εμπορίου δίνει ορθότερη πληροφόρηση για τις διακυμάνσεις των τιμών των αγαθών τι στιγμή που αυτές πρέπει να προσδιοριστούν ανεξάρτητα από τη μεγάλη διακύμανση που πιθανόν να παρατηρείται στις τιμές των προϊόντων. Έτσι οι συντελεστές μεταβλητικότητας συγκρινόμενοι μεταξύ τους δίνουν τη σωστή πληροφόρηση. Η χρησιμοποίηση της τυπικής απόκλισης είναι δυνατόν να μας οδηγήσει σε λανθασμένα συμπεράσματα λόγω των μεγάλων διαφορών που παρατηρούνται στις τιμές των διαφόρων αγαθών. Το ίδιο βέβαια ισχύει και για το μέσο αριθμητικό ο οποίος εάν υπολογίζεται μεταξύ ακραίων τιμών μειώνει την αξιοπιστία του.

4.3.5. Οι Τυποποιημένες (Ανοιγμένες) Διαφορές

Με τις τυποποιημένες διαφορές μπορούμε να συγκρίνουμε τη διακύμανση δύο ή περισσότερων μεταβλητών (ή κατανομών συχνοτήτων), ανεξάρτητα από τις μονάδες μέτρησής τους.

Χρησιμοποιώντας τον τυπικό μετασχηματισμό (βλ. επόμενο κεφάλαιο) $z_i = \frac{x_i - \bar{x}}{\sigma}$, μετατρέπουμε τις απόλυτες διαφορές των τιμών της μεταβλητής από το μέσο τους ($x_i - \bar{x}$) σε σχετικές, διαιρώντας με την τυπική τους απόκλιση. Μ' αυτόν τον τρόπο μεταφέρουμε την αρχή της μεταβλητής (x_i) από το μηδέν στο μέσο αριθμητικό της ($\bar{x}$).

Παράδειγμα 4.7

Ένας φοιτητής, βαθμολογήθηκε στις εξετάσεις του Ιουνίου 2008 στο μάθημα της Στατιστικής με 8. Ένας άλλος φοιτητής βαθμολογήθηκε στο ίδιο μάθημα στις εξετάσεις του Ιουνίου 2009 με 7. Με κριτήριο το βαθμό στις εξετάσεις, ποιος από τους δύο φοιτητές είναι καλύτερος στη Στατιστική;

Απάντηση:

Αν δε βιαστούμε να απαντήσουμε, διαπιστώνουμε ότι, ουσιαστικά, μας ζητούν να συγκρίνουμε «ανόμοια πράγματα», αφού πρέπει να συγκρίνουμε δυο τιμές η κάθε μια από τις οποίες ανήκει σε διαφορετική κατανομή. Η τιμή 8 ανήκει στην κατανομή βαθμολογίας των εξετάσεων του Ιουνίου 2008 ενώ η τιμή 7 ανήκει στην κατανομή της βαθμολογίας των εξετάσεων του Ιουνίου 2009. Για να συγκριθούν επομένως οι δύο τιμές, πρέπει να προσδιοριστεί πρώτα η σχετική απόσταση της κάθε μίας μέσα στην κατανομή της.

Έτσι, αν οι βαθμολογίες των φοιτητών τον Ιούνιο 2008 είχαν μέση τιμή 7,5 και τυπική απόκλιση 0,6 και τον Ιούνιο του 2009 είχαν μέση τιμή 5,5 και τυπική απόκλιση 1,1 τότε είναι προφανές ότι το κλάσμα $\frac{8-7{,}5}{0{,}6}=\frac{0{,}5}{0{,}6}=+0{,}8$ εκφράζει την απόσταση-απόκλιση της τιμής 8 από τη μέση τιμή της κατανομής της, σε μονάδες τυπικής απόκλισης. Δηλαδή, δείχνει «πόσες φορές χωράει η τυπική απόκλιση 0,6 στην απόσταση 8-7,5». Ομοίως, το κλάσμα $\frac{7-5{,}5}{1{,}1}=\frac{1{,}5}{1{,}1}=+1{,}4$ δείχνει «πόσες φορές χωράει η τυπική απόκλιση 1,1 στην απόσταση 7-5,5». Είναι, πλέον, φανερό ότι ο βαθμός 7 είναι καλύτερος από το βαθμό 8 με την έννοια ότι απέχει από τη μέση τιμή της κατανομής του +1,4 τυπικές αποκλίσεις ενώ ο βαθμός 8 απέχει από τη μέση τιμή της δικής του κατανομής +0,8 τυπικές αποκλίσεις. Δηλαδή, ο βαθμός 7 είναι 1,4 τυπικές αποκλίσεις μεγαλύτερος από τη μέση τιμή της κατανομής του ενώ ο βαθμός 8 είναι 0,8 τυπικές αποκλίσεις μεγαλύτερος από τη μέση τιμή της δικής του κατανομής.

Τέλος, μπορούμε να καθορίσουμε με βάση την **τυπική απόκλιση** διαστήματα γύρω από τη **μέση τιμή** στα οποία να βρίσκεται συγκεκριμένο ποσοστό παρατηρήσεων[4]. Η ανισότητα του Chebyshev μας βεβαιώνει ότι: το ποσοστό των παρατηρήσεων που βρίσκεται π.χ. στο διάστημα ($\bar{X}$-2σ, $\bar{X}$+2σ) είναι τουλάχιστον 75%. Άρα, όσο πιο «στενό» είναι αυτό το διάστημα (δηλαδή όσο πιο μικρή είναι η τυπική απόκλιση), τόσο πιο κοντά στη μέση τιμή είναι οι παρατηρήσεις και κατά συνέπεια τόσο πιο μικρή είναι η μεταβλητότητα των παρατηρήσεων. Γενικά, η ανισότητα του Chebyshev μας λέει ότι: το ποσοστό των παρατηρήσεων που βρίσκονται στο διάστημα ($\bar{X}$-κσ, $\bar{X}$+κσ) είναι τουλάχιστον $1-\frac{1}{k}$ για κάθε k>1 .

[4] Δηλαδή κάτι ανάλογο με τα διαστήματα που καθορίζουμε με βάση εκατοστημόρια. Π.χ. γνωρίζουμε ότι στο διάστημα P_{90} - P_{10} βρίσκεται το 80% των παρατηρήσεων.

Ειδική περίπτωση:

Αν η κατανομή των δεδομένων είναι κανονική τότε:

Στο διάστημα ($\bar{X}$-σ, $\bar{X}$+σ) βρίσκεται το 68% περίπου των παρατηρήσεων.

Στο διάστημα ($\bar{X}$-2σ, $\bar{X}$+2σ) βρίσκεται το 95% περίπου των παρατηρήσεων.

Στο διάστημα ($\bar{X}$-3σ, $\bar{X}$+3σ) βρίσκονται όλες σχεδόν οι παρατηρήσεις (99,7%).

4.4 Οι ροπές κατανομής συχνοτήτων

Είδαμε προηγούμενα την έννοια της διακύμανσης. Η έννοια της βασιζόταν στην ιδέα του προσδιορισμού ενός μέτρου απόκλισης από το μέσο αριθμητικό. Αυτή ακριβώς είναι η έννοια των ροπών μία από τις οποίες είναι και η διακύμανση. Η ροπή υπολογίζεται από το γινόμενο της συχνότητας της τάξης και της απόστασης από το σημείο της μεταβλητής x_1 το οποίο θεωρήσαμε σαν αρχή. Το άθροισμα όλων αυτών των γινομένων όλων των τάξεων αφού διαιρεθούν με το συνολικό αριθμό των συχνοτήτων της κατανομής, μας δίνει την πρώτη ροπή. Όταν ως αρχή λαμβάνεται ο μέσος αριθμητικός οι ροπές ονομάζονται κεντρικές και δίνονται από τους τύπους:

$\mu_1 = \dfrac{\sum f_i (x_i - \bar{x})}{N}$ (πρώτη κεντρική ροπή)	**(4.23)**

$\mu_2 = \dfrac{\sum f_i (x_i - \bar{x})^2}{N}$ (δεύτερη κεντρική ροπή)	**(4.24)**

$\mu_t = \dfrac{\sum f_i (x_i - \bar{x})^t}{N}$ (t-ιοστή κεντρική ροπή)	**(4.25)**

4.5 Οι παράμετροι ασυμμετρίας

Ο προσδιορισμός της ύπαρξης συμμετρίας ή μη μιας κατανομής θεωρείται βασικό χαρακτηριστικό της μορφής και μας βοηθά στην αξιολόγηση πολλών κατανομών οι οποίες πιθανόν να έχουν ίδια μέση τιμή, ίδια διασπορά, αλλά διαφορετικό βαθμό ασυμμετρίας. Για τη μέτρηση του βαθμού ασυμμετρίας των κατανομών συχνοτήτων χρησιμοποιούμε τις παραμέτρους ασυμμετρίας.

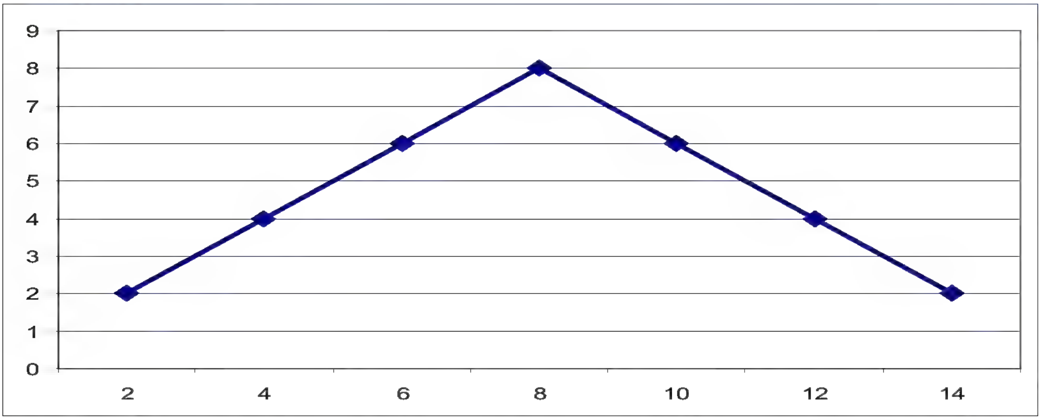
Έτσι συμμετρική είναι η κατανομή εκείνη που έχει το ίδιο πλήθος τιμών της μεταβλητής σε ίσες αποστάσεις από το Σημείο Μέγιστης Συχνότητας και από τη μέση αριθμητική τιμή της μεταβλητής.

Για παράδειγμα ας δούμε την παρακάτω κατανομή:

x_i	2	4	6	8	10	12	14
Συχνότητα f_i	2	4	6	8	6	4	2

Το σημείο μέγιστης συχνότητας των x_i είναι $M_0=8$ και οι τιμές της κατανέμονται συμμετρικά γύρω από αυτό. Η γραφική απεικόνιση της κατανομής αυτής παρουσιάζεται στο διάγραμμα 4.1.

Διάγραμμα 4.1



Για τη μέτρηση του βαθμού ασυμμετρίας μιας κατανομής έχουν προταθεί κυρίως από τους Bowley και Pearson οι παρακάτω συντελεστές ασυμμετρίας:

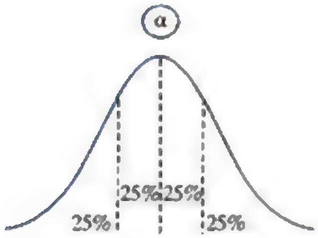
Συντελεστής **Bowley**

$$S_k(B)=\frac{(Q_3-M_e)-(M_e-Q_1)}{(Q_3-M_e)+(M_e-Q_1)}=\frac{Q_3+Q_1-2\cdot M_e}{Q_3-Q_1}$$	**(4.26)**

Αυτός ο συντελεστής ασυμμετρίας του Bowley $S_k(B)$ παίρνει τιμές μεταξύ της αρνητικής και της θετικής μονάδας.

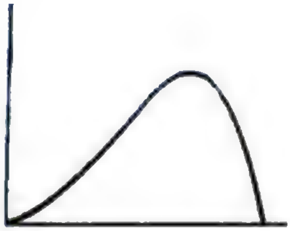
- Όταν $S_k(B)=0$ τότε, αφού $(Q_3-M_e)=(M_e-Q_1)$, η κατανομή είναι συμμετρική όπως φαίνεται στο παρακάτω Διάγραμμα 4.2.

Διάγραμμα 4.2



- Όταν $S_k(B)<0$ τότε έχουμε περίπου αρνητική ασυμμετρία της κατανομής, όπως φαίνεται στο διάγραμμα 4.3.

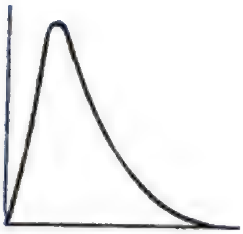
Διάγραμμα 4.3



Αρνητική ασυμμετρία

- Αντίστοιχα όταν $S_k(B)>0$, έχουμε την περίπτωση της σχεδόν θετικής ασυμμετρίας, όπως φαίνεται στο διάγραμμα 4.4.

Διάγραμμα 4.4



Θετική ασυμμετρία

Συντελεστής **Pearson**

Ο Κ. Pearson ($S_k(P)$) προτείνει, κατ' αρχάς, το συντελεστή:

$$S_k(P) = \frac{\bar{x} - M_o}{\sigma}$$	**(4.27)**

Και στην περίπτωση αυτής της σχετικής παραμέτρου (όπως και η $S_k(P)$ αφού είναι λόγοι μεγεθών της ίδιας μονάδας μέτρησης) όταν $S_k(P)=0$ έχουμε συμμετρική κατανομή ενώ θετική ή αρνητική ασυμμετρία έχουμε όταν $S_k(P)>0$ ή $S_k(P)<0$, αντίστοιχα (δηλαδή όταν αντίστοιχα έχουμε ($\bar{x}>M_0$) ή ($\bar{x}<M_0$).

Η πιο σωστή και ασφαλέστερη όμως παράμετρος ασυμμετρίας είναι αυτή του Κ. Pearson που χρησιμοποιεί τις κεντρικές ροπές (τρίτη και δεύτερη)

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$	**(4.28)**

Στις συμμετρικές κατανομές οι περιττής τάξεως κεντρικές ροπές είναι πάντοτε ίσες με μηδέν.

Έτσι παρά το γεγονός ότι ο β_1 θα είναι πάντα θετικός ή μηδέν ($\beta_1 \geq 0$) η θετική ή αρνητική ασυμμετρία μιας κατανομής θα κρίνεται από την τιμή της τρίτης κεντρικής ροπής ($\mu_3>0$ ή $\mu_3<0$ αντίστοιχα).

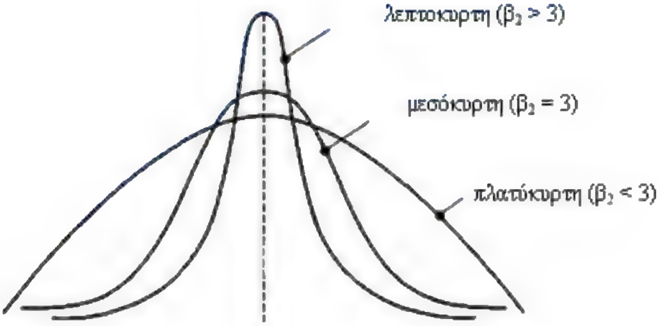
4.6 Οι παράμετροι κύρτωσης

Η κύρτωση μιας κατανομής μετράει το βαθμό συγκέντρωση των τιμών της μεταβλητής στην περιοχή των άκρων και του μέσου αριθμητικού. Για τον προσδιορισμό του βαθμού κύρτωσης χρησιμοποιούμε το συντελεστή σχετικής κύρτωσης που δίνεται από τον τύπο:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{(\sigma^2)^2}$$	**(4.29)**

Έτσι σε κατανομές με τον ίδιο μέσο αριθμητικό και την ίδια τυπική απόκλιση οι τιμές των μεταβλητών τους μπορεί να βρίσκονται κοντά στα άκρα ή το μέσο διαφοροποιώντας τη μορφή της κατανομής. Ενδεικτικό είναι το διάγραμμα 4.5 με τις καμπύλες συχνοτήτων τριών κατανομών.

Διάγραμμα 4.5



Οι κατανομές που παρουσιάζουν σχετικά μεγαλύτερη συγκέντρωση των παρατηρήσεων τους περί το μέσο ($\bar{x}$) ονομάζονται **Λεπτόκυρτες** και έχουν συντελεστή κύρτωσης $\beta_2>3$ λαμβάνοντας σαν βάση σύγκρισης τις **Μεσόκυρτες** όπως η Κανονική Κατανομή Πιθανότητας με $\beta_2=3$. Κατ' αναλογία οι κατανομές με συντελεστή κύρτωσης $\beta_2<3$ ονομάζονται **Πλατύκυρτες**.

Οι κατανομές που σε όλα τα διαστήματα τάξης τους έχουν τον ίδιο αριθμό συχνοτήτων ονομάζονται ορθογώνιες και έχουν συντελεστή κύρτωσης $1{,}7<\beta_2<1{,}8$. Όταν $\beta_2<1{,}7$ η κατανομή συχνότητας έχει μορφή σχήματος U.

4.7 Παραδείγματα

Παράδειγμα 4.8

Για τις παρακάτω παρατηρήσεις ζητείται:

α) Να ταξινομηθούν σε στατιστικούς πίνακες σχετικών συχνοτήτων και δεξιόστροφων αθροιστικών σχετικών συχνοτήτων. Να εξηγήσετε σύντομα τη χρησιμοποιούμενη μεθοδολογία σας.

β) Να υπολογίσετε και να ερμηνεύσετε για το δείγμα τις στατιστικές του μέσου αριθμητικού, της διαμέσου και της επικρατούσας τιμής (ή σημείου μεγίστης συχνότητας ή τύπου), των αταξινόμητων δεδομένων. Σχολιάστε σύντομα.

8	5	9	10	7
1	4	5	6	9
2	7	11	4	6
3	5	6	8	8

Απάντηση:

α) Για την ταξινόμηση των παρατηρήσεων χρησιμοποιείται ο εμπειρικός Κανόνας του Stugres:

δ=Εύρος/Πλήθος Τάξεων.

Εύρος (R)=Xmax-Xmin=11-1=10

Πλήθος Τάξεων (κ)=1+3,33λογ(n)=5,33≈5. Επομένως δ=10/5=2.

Στη συνέχεια κατασκευάζουμε τον πίνακα της κατανομής μας:

Τάξεις X_i	Απόλ.Συχν.(f_i)	Σχετ.Συχν.(f_i%)	Δεξ. Αθρ. Σειρά	Δεξ. Αθρ. Σχετ. Συχν. (F_i%)
[1, 3]	3	0.15	μέχρι 3	0.15
(3, 5]	6	0.3	μέχρι 5	0.45
(5, 7]	5	0.25	μέχρι 7	0.7
(7, 9]	4	0.2	μέχρι 9	0.9
(9, 11]	2	0.1	μέχρι 11	1
Σύνολο	20	1		

β) Ιδιαίτερη έμφαση θα δοθεί στην ερμηνεία των τιμών των ζητούμενων στατιστικών.

Μέσος Αριθμητικός: $\overline{X} = \frac{\sum X_i}{n} = 6{,}1$

Διάμεσος: Η παρατήρηση 6,0 στη θέση ($\frac{n+1}{2}$) των 20 κατ' αύξουσα διάταξη ταξινομημένων παρατηρήσεων του δείγματος.

Επικρ. Τιμή: 5 η συχνότερα εμφανιζόμενη παρατήρηση του δείγματος.

Από τον υπολογισμό των παραπάνω παραμέτρων συμπεραίνουμε ότι υπάρχει μέτρια θετική ασυμμετρία.

Παράδειγμα 4.9

Οι πωλήσεις (σε εκατ. €) του τελευταίου μήνα για κάθε μία από τις 15 θυγατρικές της εταιρίας ALIMENTI SA ήταν: 0.1, 0.1, 0.25, 0.25, 0.25, 0.35, 0.4, 0.53, 0.9, 1.25, 1.35, 2.45, 2.71, 3.09 και 4.1. Για το δείγμα αυτό ζητείται:

1. Να υπολογίσετε το μέσο αριθμητικό, τη διάμεσο και το σημείο μέγιστης συχνότητας.

2. Να ερμηνεύσετε οικονομικά τις τιμές των στατιστικών που υπολογίσατε.

3. Με βάση την απάντηση στο 1ο ερώτημα να περιγράψετε την ασυμμετρία αυτής της κατανομής.

Απάντηση:

1. Υπολογίζουμε το μέσο αριθμητικό $\overline{X} = \sum X_i / n = 18.08/15 = 1.21$, ,

$Me = X_{(n/2)+(1/2)} = X_{(15/2)+(1/2)} = X_8 = 0.53$, $M_0 = 0.25$

2. **Μέσος Αριθμητικός**: εάν όλες οι θυγατρικές έκαναν τις ίδιες πωλήσεις τότε αυτές θα είχαν την τιμή του μέσου αριθμητικού. **Διάμεσος**: 7 θυγατρικές έκαναν πωλήσεις κάτω από 530,000€ και άλλες 7 πάνω από το ποσό αυτό. Τύπος: το συχνότερα εμφανιζόμενο ύψος πωλήσεων είναι 250,000€.

3. Αφού $\overline{X} > M_e > M_o$ η κατανομή είναι θετικά ασυμμετρική ή ασύμμετρη προς τα δεξιά.

Παράδειγμα 4.10

Ο χρόνος (σε λεπτά) που απαιτείται για να ολοκληρώσουν την ηλεκτρονική επεξεργασία των φορολογικών δηλώσεων του 2008 οι 30 εργαζόμενοι ενός τμήματος του ΚΕΠΥΟ είναι: 10, 16, 15, 9, 12, 11, 14, 12, 18, 15, 10, 16, 15, 14, 9, 11, 17, 12, 13, 11, 14, 13, 16, 14, 17, 13, 14, 11, 12, 15. Για τις παρατηρήσεις αυτές ζητείται:

1. Να τις ταξινομήσετε σε πέντε τάξεις ίσου πλάτους (Κατανομή απόλυτων και σχετικών συχνοτήτων).

2. Να περιγράψετε τη μεταβλητικότητά τους, χρησιμοποιώντας τις τιμές της τυπικής απόκλισης και των τεταρτημορίων Q3 και Q1.

Απάντηση:

1. Κατασκευάζουμε το πίνακα της κατανομής μας

X	f_i	f_i%	X_i (Πρ.Μ.)	f_i x X_i	fi x (X_i-X(avg))^2	F_i
[9, 11)	4	0,13	10	40	57,76	4
[11, 13)	8	0,27	12	96	25,92	12
[13, 15)	8	0,27	14	112	0,32	20
[15, 17)	7	0,23	16	112	33,88	27
[17, 19)	3	0,10	18	54	52,92	30
Σύνολο	30	1,00		414	171	

Το πλάτος και το πλήθος των τάξεων υπολογίζεται με τον εμπειρικό κανόνα του **Sturges**: δ=(Xmax-Xmin)/5=(18-9)/5≈2

2. $\bar{X}$ =(ΣfiXi)/Σfi=414/30=13.80,
s=((Σfi(Xi-)^2)/n-1)^(½)=171/29=2.43
ή χωρίς βαθμό ελευθερίας 2.38

$Q1=X_{Q1}+(δ/f_{Q1})(n/4\text{-}Fi\text{-}1)=11+(2/8)(7.5-4)=11.88$
$Q3=X_{Q3}+(δ/f_{Q3})(3n/4\text{-}_{Fi-1})=15+(2/7)(22.5-20)=15.71$

3. Η μεταβλητότητα του χρόνου απασχόλησης του 50% των εργαζομένων (15) είναι 3.84 (R=Q3-Q1=15.71-11.88=3.84), με μέγιστο χρόνο του 75% τα 15.71 (Q3) και ελάχιστο χρόνο του 25% τα 11.88’ (Q1).

Παράδειγμα 4.11

1. Η εταιρία ENRON Α.Ε. χορήγησε σε 5000 στελέχη της διάφορα δάνεια για αγορά κατοικίας. Σε μια μελέτη που έγινε, ελήφθη δείγμα τριακοσίων δανείων το οποίο παρουσιάζεται στον παρακάτω πίνακα κατανομής. Να περιγράψετε στατιστικά το δείγμα με τη χρήση του μέσου αριθμητικού, της τυπικής απόκλισης, του συντελεστή μεταβλητικότητας, της διαμέσου και του συντελεστή ασυμμετρίας (Pearson).

2. Είναι δυνατόν το εξήντα τοις εκατό των στελεχών της ENRON A.E. να έχει πάρει δάνειο πάνω από τον μέσο αριθμητικό όρο;

Κατανομή Δανείων	
Σύνορα Δανείων	Αριθμός Πελατών
[0, 10)	49
[10, 20)	90
[20, 30)	78
[30, 40)	42
[40, 50)	26
[50, 60)	15

Σημ.: Ποσά σε εκατ. €

Απάντηση:

1.

Τάξεις	Απολ.Συχν. f_i	Κεντρ.Όροι X_i	$f_i \times X_i$	$(X_i - \overline{X})^2$	$fi \times (X_i - \overline{X})^2$
[0, 10)	49	5	245	337,35	16529,9
[10, 20)	90	15	1350	70,01	6300,6
[20, 30)	78	25	1950	2,67	208,0
[30, 40)	42	35	1470	135,33	5683,7
[40, 50)	26	45	1170	467,99	12167,65
[50, 60)	15	55	825	1000,65	15009,7
Σύνολο	*300*		*7010*		*55899,67*

Τάξεις	Απόλυτη Συχνότητα f_i	Δεξιόστροφη Αθροιστική Σειρά	Δεξιόστροφη Αθρ/κή Απόλυτη Συχνότητα F_i
[0, 10)	49	έως 10	49
[10, 20)	90	έως 20	139
[20, 30)	78	έως 30	217
[30, 40)	42	έως 40	259
[40, 50)	26	έως 50	285
[50, 60)	15	έως 60	300
Σύνολο	*300*		

Μέσος Αριθμητικός: $\bar{X} = \frac{\sum f_i X_i}{n} = \frac{7010}{300}$ =**23,367**	*Τυπική Απόκλιση:* $s = \sqrt{\frac{\sum f_i (X_i - \bar{X})^2}{\sum f_i}} = \sqrt{\frac{55899,67}{300}} = \sqrt{186,3}$ =**13,65**
Συντελεστής Μεταβλητητας: CV= s / $\bar{X}$ =13,65/23,367 =**0,58 ή 58%**	*Διάμεσος:* $M_e = a_{Me} + \frac{\delta}{f_{Me}}(\frac{n}{2} - F_{i-1}) = 20 + \frac{10}{78}\left(\frac{300}{2} - 139\right)$ =**21,41**
Επικρατούσα Τιμή: $M_o = a_{Mo} + \delta(\frac{\Delta_1}{\Delta_1 + \Delta_2}) = 10 + 10\left(\frac{90-49}{(90-49)+(90-78)}\right)$ =**17,74**	*Συντελεστής Ασυμμετρίας (Pearson):* $S_K(P) = \frac{\bar{X} - M_O}{s} = \frac{23,367 - 17,74}{13,65}$ =**0,41 ή 41%**

Τα 150 δάνεια ήταν πάνω από 21,41 εκ. € και τα υπόλοιπα κάτω από το ύψος αυτό.

Το ύψος των δανείων που χορηγεί η Τράπεζα κυμαίνεται κατά μέσο όρο από 9,7 (=23,367-13,65) εκ. € έως 37 (=23,367+13,65) εκ. €.

Η σχετική διασπορά των χορηγουμένων δανείων είναι υψηλή (58%).

Η κατανομή των δανείων παρουσιάζει ελαφρά θετική ασυμμετρία (41%).

2. $\bar{X}$ (=23,367) > M_e (=21,41)

Αφού πάνω από τη διάμεση τιμή (M_e) έχουν δοθεί 150 δάνεια, δεν είναι δυνατόν πάνω από το μέσο αριθμητικό όρο ($\bar{X}$), για τον οποίο βρήκαμε $\bar{X}$ > M_e, να έχουν δοθεί περισσότερα (0,60x300=180) δάνεια.

5. Εκπαιδευτική Ενότητα

- **Στοιχεία Πιθανοθεωρίας Ι**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατανοεί τα παρακάτω στοιχειά θεωρίας Πιθανοτήτων: Τυχαίο πείραμα, Ενδεχόμενα, δειγματικός χώρος.
- Να κατανοεί την έννοια Πιθανότητα και τις βασικές ιδιότητες των πιθανοτήτων.
- Να αντιλαμβάνεται τα σχετικά με τις τυχαίες μεταβλητές και τις κατανομές πιθανότητας: έννοιες, κατανομές, παράμετροι και ιδιότητες τους.
- Να αναγνωρίζει και να κατανοεί τις σχέσεις μεταξύ τυχαίων μεταβλητών.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Τυχαίο πείραμα
- Ενδεχόμενα
- Δειγματικός χώρος
- Κατανομές
- Παράμετροι κατανομών

5.1 Εισαγωγή

Στα προηγούμενα κεφάλαια εξετάσαμε τα στάδια ανάλυσης της περιγραφικής στατιστικής. Στα τελευταία τα συμπεράσματά μας αφορούν μόνο τα στοιχεία του δείγματος που μελετάμε. Είδαμε ότι σε πρώτη φάση, με βάση τους πίνακες και τα διαγράμματα, συμπυκνώνουμε τη διαθέσιμη στατιστική πληροφορία, και μετά σε δεύτερη φάση, η σύμπτυξη αυτή γίνεται ακόμα μεγαλύτερη, με την ακρίβεια για τη μορφολογία του δείγματος, που μας δίνουν οι στατιστικές που υπολογίζουμε, δηλαδή τα μέτρα τάσης, θέσης, διασποράς, ασυμμετρίας και κύρτωσης.

Εντούτοις, τελικός στόχος της στατιστικής ανάλυσης είναι η εξαγωγή αξιόπιστων συμπερασμάτων για το γεννήτορα πληθυσμό του δείγματος που μελετάμε ή εναλλακτικά για τη θεωρία σύμφωνα με την οποία πιστεύουμε ότι παρήχθησαν οι παρατηρήσεις του δείγματός μας. Με άλλα λόγια, στόχος της στατιστικής ανάλυσης είναι η στατιστική επαγωγή (statistical inference), δηλαδή η γενίκευση των συμπερασμάτων της περιγραφικής ανάλυσης στον πληθυσμό από τον οποίο προέρχεται το τυχαίο δείγμα που μελετήσαμε. Κάτω από ορισμένες προϋποθέσεις, που θα ξεκαθαρίσουμε παρακάτω, οι «δειγματικές τιμές», π.χ., για το μέσο αριθμητικό ($\bar{X}$) ή την τυπική απόκλιση (s) ή τις άλλες στατιστικές, μπορούν να γενικευθούν έτσι ώστε να κάνουμε «εκτιμήσεις» για τις αντίστοιχες πληθυσμιακές «παραμέτρους», άγνωστες μεν αλλά αληθινές, και για τις οποίες κυρίως ενδιαφερόμαστε (εδώ αντίστοιχα μ και σ).

Η γέφυρα που ενώνει την περιγραφική με την επαγωγική στατιστική είναι η θεωρία των πιθανοτήτων. Στοιχεία της θεωρίας αυτής θα αναφερθούν σε αυτό και στο επόμενο κεφάλαιο.

Στο παρόν κεφάλαιο θα μας απασχολήσουν κυρίως τα τρία (3) βασικά είδη ερωτημάτων που αφορούν την ανάλυση των πιθανοτήτων:

1. Τι εννοούμε όταν λέμε η «ότι πιθανότητα επιτυχίας στην υπόψη εξέταση είναι 80%»; Πώς τεκμηριώνεται η πρόταση «η πιθανότητα επιτυχούς εισαγωγής του νέου προϊόντος στην αγορά είναι 60%».

2. Πώς προσδιορίζουμε τις τιμές των πιθανοτήτων που αποδίδουμε σε ορισμένα ενδεχόμενα να συμβούν στην καθημερινή πραγματική ζωή;

3. Ποιοι είναι οι κανόνες που διέπουν τις πιθανότητες;

Το πρώτο στάδιο στην προσπάθειά μας να προβλέψουμε τι είναι πιθανόν να συμβεί είναι να μπορούμε να υπολογίσουμε τι είναι δυνατόν να συμβεί, με δεδομένο ότι τα περισσότερα φαινόμενα της πραγματικής, και ειδικά της οικονομικής ζωής χαρακτηρίζονται από αβεβαιότητα.

Στο επόμενο τμήμα εξηγούμε σύντομα την έννοια του δειγματικού χώρου και των επιμέρους του ενδεχομένων, τα οποία συνιστούν αυτό που οι στατιστικολόγοι χρησιμοποιούν για να υπολογίσουν «τι μπορεί να συμβεί». Στη συνέχεια θα αναφερθούμε σύντομα στη μεθοδολογία υπολογισμού του δειγματικού χώρου. Τα δύο πρώτα από τα παραπάνω ερωτήματα στην ανάλυση των πιθανοτήτων θα απαντηθούν σύντομα στα τμήματα 4 και 5, ενώ οι βασικές ιδιότητες των πιθανοτήτων θα παρουσιαστούν στην τελευταία παράγραφο του κεφαλαίου.

5.2 Πείραμα τύχης, ενδεχόμενα, δειγματικός χώρος

Πείραμα (experiment) ορίζεται η διαδικασία η οποία μπορεί να επαναλαμβάνεται κάτω από τις ίδιες κάθε φορά συνθήκες και να οδηγεί σε κάποιο από ορισμένα δυνατά αποτελέσματα. Αποτέλεσμα (outcome or event) ενός πειράματος είναι μια παρατήρηση ή μια μέτρησή του. Για παράδειγμα η ρίψη ενός νομίσματος n φορές στο ίδιο περιβάλλον αποτελεί ένα πείραμα, όπως επίσης οι ιατρικές έρευνες για την επίδραση των φαρμάκων στη θεραπεία των ασθενειών.

Όταν τα αποτελέσματα ενός πειράματος δεν είναι γνωστά ή δεν μπορούν να προβλεφθούν με βεβαιότητα τότε μιλάμε για πείραμα τύχης.

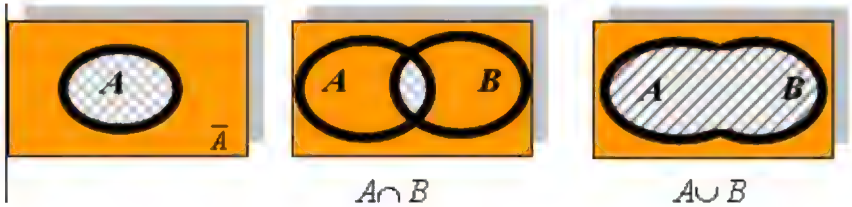
Τα επιμέρους αποτελέσματα ενός πειράματος τύχης ορίζονται ως απλά ή στοιχειώδη ενδεχόμενα ή γεγονότα (basic outcome or elementary event). Τα ενδεχόμενα συμβολίζονται με κεφαλαία γράμματα, A, B, Γ, … . Για παράδειγμα, απλά ενδεχόμενα είναι για το πείραμα τύχης, της ρίψης ενός ζαριού, κάθε ένα από τα στοιχεία του συνόλου των αριθμών Ω={1, 2, 3, 4, 5, 6}, ενώ στην περίπτωση της γέννησης 1 παιδιού, S={A, K} όπου A=αγόρι, K=κορίτσι, καθώς επίσης, στην περίπτωση της γέννησης 2 παιδιών Z={AA, AK, KA, KK}.

Τα παραπάνω τρία σύνολα Ω, S και Z ορίζουν αυτό που ονομάζεται δειγματικός χώρος (sample space) και ο οποίος περιλαμβάνει όλα τα δυνατά ενδεχόμενα ενός πειράματος τύχης. Οι δειγματικοί χώροι διακρίνονται σε συνεχείς και διακριτούς, ανάλογα αν δεν είναι ή αν είναι, αντίστοιχα, πεπερασμένοι ή με αριθμήσιμο πλήθος απλών ενδεχόμενων.

Εναλλακτικά, μπορεί να μας ενδιαφέρει συλλογή στοιχειωδών ενδεχομένων, οπότε τότε μιλάμε για «σύνθετα ενδεχόμενα» ή απλώς «ενδεχόμενα». Προφανώς το σύνολο των αποτελεσμάτων ενός πειράματος τύχης μπορεί να θεωρηθεί ένα ενδεχόμενο.

Τα διαγράμματα Venn (βλ. Σχήμα 5.1) είναι σχηματικές παρουσιάσεις συνόλων και χρησιμοποιούνται για να δείξουν τις σχέσεις μεταξύ ενδεχόμενων.

Σχήμα 5.1 Διαγράμματα Venn: σχέσεις ενδεχομένων



Ένωση δύο ενδεχόμενων A και B του δειγματικού χώρου S είναι το ενδεχόμενο έστω $Γ=A\cup B$ που περιλαμβάνει τα διακεκριμένα αποτελέσματα που ανήκουν είτε στο A είτε στο B είτε και στα δύο (βλ. Σχήμα 5.1).
Τομή δύο ενδεχόμενων A και B του δειγματικού χώρου S είναι το ενδεχόμενο έστω $Δ=A\cap B$ που περιλαμβάνει τα αποτελέσματα του πειράματος τύχης που είναι κοινά στο A και στο B (βλ. Σχήμα 5.1). Όταν $A\cap B=\varnothing$ τότε τα δύο αυτά ενδεχόμενα ονομάζονται ξένα ή ασυμβίβαστα.
Διαφορά δύο ενδεχόμενων A και B του δειγματικού χώρου S είναι το ενδεχόμενο έστω E=A-B που περιλαμβάνει τα αποτελέσματα που ανήκουν στο A και όχι στο B.

Συμπλήρωμα ή αντίθετο του A ενδεχομένου του δειγματικού χώρου S ονομάζεται το ενδεχόμενο $\bar{A}$ που περιλαμβάνει τα αποτελέσματα που δεν περιέχονται στο A αλλά ανήκουν στο S (βλ. Σχήμα 5.1).

Παράδειγμα 5.1

Δύο φίλοι συναγωνίζονται σε ένα παιχνίδι τύχης για 2 διαφορετικά βραβεία. Ποια είναι τα δυνατά ενδεχόμενα του παιχνιδιού;

Απάντηση:

Το σύνολο όλων των δυνατών αποτελεσμάτων σε αυτό το πείραμα τύχης, το οποίο συνιστά και το «δειγματικό του χώρο» είναι:

S= {(0, 0), (0, 1), (1, 0), (1, 1), (2, 0), (0, 2)}

Όπου
(1, 1) σημαίνει το απλό ενδεχόμενο κάθε ένας από τους δύο φίλους να κερδίσει από ένα βραβείο, ενώ (2, 0) εκφράζει το ενδεχόμενο ο πρώτος από τους δύο να κερδίσει και τα δύο βραβεία, κ.ο.κ.

Σύνθετα ενδεχόμενα μπορεί να είναι οι συλλογές:

A= { (0, 1), (1, 0) },
που εκφράζει το σύνθετο ενδεχόμενο κάθε ένας από τους δύο φίλους να παίρνει ένα βραβείο.

B= { (0, 0), (0, 1), (1, 0), (1, 1) }
που εκφράζει το σύνθετο ενδεχόμενο κάθε ένας από τους δύο φίλους να παίρνει είτε κανένα είτε ένα βραβείο (ισοδύναμα, κάθε ένας παίρνει το πολύ ένα βραβείο).

Γ = { (1, 1) }
που εκφράζει το σύνθετο ενδεχόμενο και οι δύο φίλοι να παίρνουν από ένα βραβείο.

Κάποιο από τα σύνθετα ενδεχόμενα A, B, Γ λέμε ότι έχει συμβεί εάν οποιοδήποτε από τα στοιχειώδη αποτελέσματα που αυτό περιέχει έχει συμβεί.

Αδύνατο ενδεχόμενο είναι εκείνο που δεν περιέχει κανένα αποτέλεσμα του πειράματος τύχης. Έτσι, για το παράδειγμά μας το A είναι αδύνατο (A=∅) εάν στο υπόψη παιχνίδι τύχης και τα δύο βραβεία τα πήρε ο πρώτος από τους δύο φίλους, δηλαδή είχαμε το ενδεχόμενο (2, 0).

Ανεξάρτητα ενδεχόμενα είναι εκείνα που η εμφάνιση ενός δεν αποκλείει την εμφάνιση του άλλου. Στο παράδειγμά μας τα στοιχειώδη ενδεχόμενα του A είναι ανεξάρτητα μεταξύ τους.

Αμοιβαία αποκλειόμενα ή ξένα λέγονται δύο ενδεχόμενα που η έλευση του ενός αποκλείει την εμφάνιση του άλλου. Στο παράδειγμά μας τα στοιχειώδη ενδεχόμενα π.χ. (2, 0) και (0, 1) είναι αμοιβαία αποκλειόμενα.

Βέβαια ενδεχόμενα είναι αυτά που θα συμβούν οπωσδήποτε κατά την εκτέλεση του πειράματος τύχης. Στο παράδειγμά μας τα στοιχειώδη ενδεχόμενα του δειγματικού χώρου S είναι βέβαια, με την έννοια ότι κάποιο από αυτά οπωσδήποτε θα συμβεί, δηλ. το S είναι ένα βέβαιο ενδεχόμενο.

Πιθανό ενδεχόμενο είναι εκείνο που η εμφάνισή του δεν είναι βέβαιη. Στο παράδειγμά μας το ενδεχόμενο (0, 2) δηλαδή και τα δύο βραβεία να πάρει ο δεύτερος από τους δύο φίλους είναι πιθανό ενδεχόμενο.

5.3 Προσδιορισμός δειγματικού χώρου

Η καταγραφή των ενδεχόμενων δειγματικού χώρου σε απλά, χωρίς πολλές επαναλήψεις, πειράματα τύχης διευκολύνεται ιδιαίτερα με τη χρήση των **δενδροδιαγραμμάτων** (tree diagram) και των **πινάκων συνάφειας** (contingency tables).

Το επόμενο παράδειγμα πιστεύεται ότι θα βοηθήσει στην κατανόηση του τρόπου λειτουργίας των δενδροδιαγραμμάτων στη συγκεκριμένη περίπτωση.

Παράδειγμα 5.2

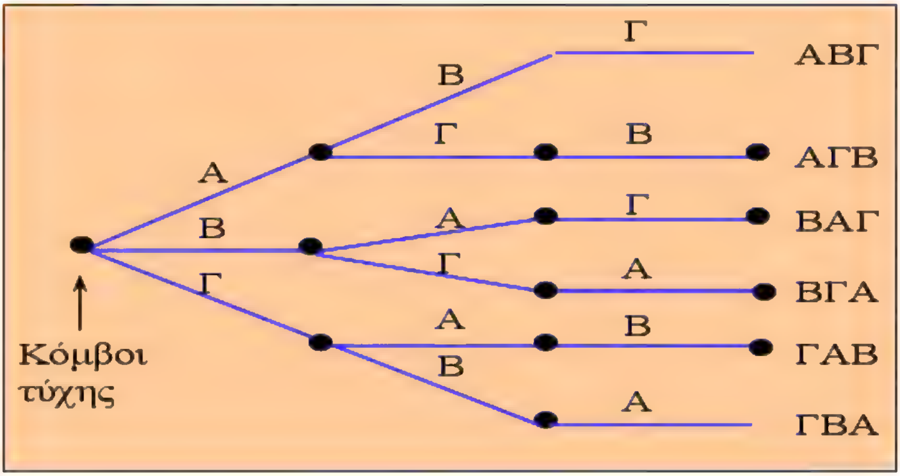
Να καταγράψετε τα δυνατά αποτελέσματα (δειγματικό χώρο) πειράματος τύχης το οποίο συνίσταται στην επιλογή από καλάθι που περιέχει 3 σφαίρες με τα γράμματα Α, Β, Γ. Με άλλα λόγια ζητείται το σύνολο των «**διατάξεων**» των τριών γραμμάτων. Μας ενδιαφέρει δηλ. η σειρά των γραμμάτων.

Απάντηση:

Όπως φαίνεται στον παρακάτω Σχήμα 5.2 μπορεί εύκολα να διαπιστωθεί ότι τα στοιχειώδη ενδεχόμενα είναι συνολικά 6. Πιο συγκεκριμένα ο δειγματικός χώρος θα είναι:

S={ΑΒΓ, ΑΓΒ, ΒΑΓ, ΒΓΑ, ΓΑΒ, ΓΒΑ}

Σχήμα 5.2 Δενδροδιάγραμμα



Γενικότερα αν ένα πείραμα τύχης μπορεί να εκτελεστεί κατά n τρόπους καθένας από τους οποίους μπορεί επίσης να εκτελεστεί κατά m τρόπους τότε το σύνθετο αποτέλεσμα μπορεί να γίνει, σύμφωνα με την πολλαπλασιαστική αρχή κατά n x m τρόπους. Η πολλαπλασιαστική αρχή δίνει τις δυνατές διατάξεις δειγματικού χώρου.

Παράδειγμα 5.3

Ακτοπλοϊκή εταιρεία διαθέτει 4 πλοία και 6 καπετάνιους για τα δρομολόγια που έχει αναλάβει. Να δείξετε το σύνολο των δυνατών τρόπων (δειγματικός χώρος), να αντιστοιχηθούν οι καπετάνιοι στα πλοία.

Απάντηση:

Δεδομένου ότι n=4 και m=6 ο ζητούμενος δειγματικός χώρος θα έχει 24 στοιχεία, δηλ.

S={11, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25, 26, 31, 32, 33, 34, 35, 36, 41, 42, 43, 44, 45, 46}

Όπου το πρώτο ψηφίο του S αφορά στο πλοίο και το δεύτερο στον καπετάνιο. Αν και με πολλά, εντούτοις πεπερασμένα στοιχεία, το πρόβλημα αυτό θα μπορεί να λυθεί και με τη χρήση δενδροδιαγραμμάτων.

Τα παραπάνω ισχύουν για δειγματοληψία χωρίς επανάθεση. Αν όμως, γίνεται επανάθεση, τότε ο γενικός κανόνας για το πλήθος των αποτελεσμάτων δειγματικού χώρου S, σε n επαναλήψεις π.τ., αμοιβαία αποκλειόμενων m αποτελεσμάτων που τον εξαντλούν είναι m^n. Έτσι, το πλήθος των αποτελεσμάτων του δειγματικού δίνεται από τη σχέση:

m^n	**(5.1)**

Σε προβλήματα που θέλουμε το πλήθος των δυνατών ενδεχομένων, χωρίς όμως να μας ενδιαφέρουν τα ακριβή στοιχεία τους, χρησιμοποιούμε δύο πολύ σημαντικές σχέσεις που αναφέρονται στις **διατάξεις** (permutations) και στους **συνδυασμούς** (combinations).

Ο κανόνας με τον οποίο σχηματίζονται τα ενδεχόμενα στην περίπτωση των διατάξεων είναι: δύο οποιαδήποτε ενδεχόμενα να διαφέρουν μεταξύ τους είτε κατά κάποιο στοιχείο, είτε κατά τη διάταξη των στοιχείων, είτε και κατά τα δύο αυτά.
Η παρακάτω σχέση (5.1) δίνει το πλήθος των διατάξεων

$P(n,m)=\dfrac{n!}{(n-m)!}$	**(5.2)**

Όπου n! n παραγοντικό, n!= n(n-1)(n-2)…2·1.

Αντίθετα, αν ενδιαφερόμαστε μόνο για ομάδες στοιχείων που διαφέρουν μεταξύ τους κατά κάποιο στοιχείο, τότε μιλάμε για συνδυασμούς, ο ορισμός των οποίων δίνεται από την παρακάτω σχέση (5.3):

$C(n,m)=\dfrac{n!}{m!(n-m)!}$	**(5.3)**

Παράδειγμα 5.4

Από το σύνολο των 13 μελών εκπαιδευτικού προσωπικού ακαδημαϊκής μονάδας ο Πρόεδρος της πρέπει να επιλέξει 3 τυχαία για τη σύσταση κάποιας επιτροπής. Να δείξετε το σύνολο των δυνατών τρόπων (δειγματικός χώρος) επιλογής των 3 διδασκόντων: α) εάν λαμβάνεται υπόψη η σειρά τους και β) ανεξαρτήτως σειράς, αρκεί μόνο οι 3 διδάσκοντες να είναι διαφορετικά πρόσωπα.

Απάντηση:

Είναι προφανές ότι το **α)** ζητούμενο αφορά διατάξεις, ενώ το **β)** συνδυασμούς. Αριθμώντας τυχαία τους 13 διδάσκοντες και τοποθετώντας τα χαρτάκια με τους αριθμούς σε ένα κουτί επιλέγουμε κάποιο και συνεχίζουμε στο επόμενο χωρίς να επανατοποθετούμε στο κουτί το πρώτο (χωρίς αντικατάσταση).

α) Εδώ έχουμε την περίπτωση των διατάξεων 13 στοιχείων ανά 3, δηλ.:

$P(13,3)=\dfrac{13!}{(13-3)!}=1.716$ εναλλακτικές διατάξεις 3 διδασκόντων όπου η σειρά τους μας ενδιαφέρει για την επιτροπή, εάν πχ. ο πρώτος επιλεγόμενος είναι ο Πρόεδρος, ο δεύτερος θα είναι ο Αντιπρόεδρος κ.λπ.).

β) Στην περίπτωση των συνδυασμών των 13 στοιχείων ανά 3, θα έχουμε:

$C(13,3)=\dfrac{13!}{3!(13-3)!}=286$ εναλλακτικοί συνδυασμοί 3 διαφορετικών διδασκόντων για την επιτροπή.

Εναλλακτικά, πολύ συνηθισμένος τρόπος, επιχειρησιακά, για την καταγραφή του δειγματικού χώρου, σχετικά απλών τυχαίων φαινομένων είναι οι «Πίνακες Συνάφειας».

Παράδειγμα 5.5

Στον παρακάτω πίνακα συνάφειας δίνονται τα αποτελέσματα από δείγμα 500 φοιτητών Τμήματος Α.Ε.Ι. σχετικά με τη συμπεριφορά τους για Προγράμματα Μεταπτυχιακών Σπουδών (ΠΜΣ). **α)** Ποιος είναι ο δειγματικός χώρος του υπόψη τυχαίου πειράματος; **β)** Δώστε παραδείγματα απλών και σύνθετων ενδεχομένων.

Πίνακας 5.1 Συμπεριφορά φοιτητών σχετικά με ΠΜΣ

	Συνεχίζουν σε ΠΜΣ		
Σχεδιάζουν για ΠΜΣ	ΝΑΙ	ΌΧΙ	Σύνολο
ΝΑΙ	100	25	125
ΌΧΙ	50	325	375
Σύνολο	150	350	500

Απάντηση:

α) Ο δειγματικός χώρος είναι οι 500 φοιτητές που ανταποκρίθηκαν στις ερωτήσεις των δειγματοληπτών.

β)
- Απλά ενδεχόμενα είναι «ΝΑΙ, σχεδιάζω να παρακολουθήσω ΠΜΣ» δηλ. 125 φοιτητές, «ΟΧΙ, δεν σχεδιάζω να παρακολουθήσω ΠΜΣ» δηλ. 375 φοιτητές, «ΝΑΙ, συνεχίζω σε ΠΜΣ» δηλ. 150 φοιτητές, «ΟΧΙ, δεν συνεχίζω σε ΠΜΣ» δηλ. 350 φοιτητές.

- Το συμπλήρωμα του ενδεχομένου «ΝΑΙ, σχεδιάζω να παρακολουθήσω ΠΜΣ» είναι το ενδεχόμενο «ΟΧΙ, δεν σχεδιάζω να παρακολουθήσω ΠΜΣ».

- Σύνθετο ή συνδυασμένο ενδεχόμενο π.χ. είναι το «ΝΑΙ, σχεδιάζω να παρακολουθήσω ΠΜΣ και ΝΑΙ, συνεχίζω σε ΠΜΣ, σύμφωνα με το σχεδιασμό μου» δηλ. 100 φοιτητές.

5.4 Έννοιες, ιδιότητες και προσδιορισμός πιθανοτήτων

Δυστυχώς, για το θεμέλιο της στατιστικής επαγωγής δεν υπάρχει μόνο ένας ορισμός της πιθανότητας, για λόγους που ξεφεύγουν από τα όρια της στατιστικής επιστήμης. Θα παρουσιάσουμε σύντομα τους τρεις (3) πιο γνωστούς ορισμούς-ερμηνείες της απλής ή οριακής ή περιθώριας πιθανότητας (simple or marginal probability):

1) Κλασική ή μαθηματική πιθανότητα.
2) Εμπειρική ή στατιστική πιθανότητα.
3) Υποκειμενική πιθανότητα.

Ο κλασικός ορισμός διατυπώθηκε από τον P.S. Laplace και ονομάζεται επίσης θεωρητική ή μαθηματική ή εκ των προτέρων πιθανότητα. Αφορά στα δυνατά και αμοιβαία αποκλειόμενα ενδεχόμενα δειγματικού χώρου με το σημαντικό χαρακτηριστικό ότι είναι όλα ισοπίθανα (equally likely).

$P(A)=\frac{n(A)}{n(S)}$	**(5.4)**

Όπου:
n(A) το πλήθος των ισοπίθανων τρόπων εμφάνισης του ενδεχομένου Α (ευνοϊκές περιπτώσεις) και
n(S) το πλήθος όλων των ισοπίθανων ενδεχομένων του δειγματικού χώρου S.

Παράδειγμα 5.6

Μια κάλπη περιέχει 6 άσπρα, 2 κόκκινα και 12 πράσινα σφαιρίδια. Ανακατεύοντας τα σφαιρίδια παίρνουμε 1 κατά τυχαίο τρόπο. Ποια είναι η πιθανότητα να εξαχθεί κόκκινο σφαιρίδιο;

Απάντηση

Οι ευνοϊκές περιπτώσεις εδώ είναι τα 2 κόκκινα σφαιρίδια. Επομένως εάν ορίσουμε Α το ενδεχόμενο να πάρουμε κόκκινο σφαιρίδιο τότε n(A) = 2. Ο δειγματικός χώρος S είναι το σύνολο των σφαιριδίων δηλ. n(S) = 20 σφαιρίδια. Επομένως η ζητούμενη πιθανότητα του Α θα είναι:

$$P(A)=\frac{n(A)}{n(S)}=\frac{2}{20}=0{,}10\,.$$

Ο R. Von Mises είναι ο βασικός εκπρόσωπος της σχολής της εμπειρικής ή στατιστικής ή εκ των υστέρων πιθανότητας.

Έτσι, η εμπειρική πιθανότητα του ενδεχομένου Α εκτιμάται ως το όριο της σχετικής συχνότητας εμφάνισης του Α όταν ο αριθμός επαναλήψεων, κάτω από τις ίδιες πάντοτε συνθήκες, του τυχαίου πειράματος (τ.π.) είναι πάρα πολύ μεγάλος (τείνει στο άπειρο).

Συμβολικά μπορούμε να γράψουμε:

$P(A)=\lim_{N\to\infty}\frac{n(A)}{N},\ 0\le n(A)\le N$	**(5.5)**

Όπου
Ν το πλήθος επαναλήψεων του τυχαίου πειράματος.

Η διαφορά μεταξύ της μαθηματικής και της στατιστικής έννοιας της πιθανότητας είναι σημαντική. Για παράδειγμα, στο τυχαίο πείραμα της ρίψης ενός νομίσματος μια φορά, η πιθανότητα του ενδεχομένου Α = εμφάνιση της επιφάνειας κορώνα είναι πρώτον, κατά το μαθηματικό ορισμό P(A) = ½ αφού από τα δύο ισοπίθανα ενδεχόμενα ενδιαφερόμαστε για το ένα από αυτά. Δεύτερον, κατά τη στατιστική ερμηνεία της πιθανότητας του υπόψη

ενδεχομένου Α θα προσεγγίζουμε την P(A) = ½ όσο αυξάνουν οι επαναλήψεις του πειράματος. Έτσι, στις πρώτες 50 ρίψεις η σχετική συχνότητα εμφάνισης κορώνας μπορεί να είναι πχ. 0,59 ενώ στις 100 ρίψεις να γίνει 0,55, στις 300 να πέσει ακόμα περισσότερο έστω 0,52 και στις 500 να τείνει στο ½ έστω 0,502.

Παράδειγμα 5.7

Σε 5.000 καταγεγραμμένες περιπτώσεις εξέτασης στο μάθημα της Στατιστικής Επιχειρήσεων Ι, κάποιου οικονομικού τμήματος Α.Ε.Ι. 1.975 φοιτητές πήραν βαθμολογία μέχρι 7, ενώ 1.230 από 7 και πάνω. Ποια είναι η εμπειρική πιθανότητα του ενδεχομένου επιτυχίας στη στατιστική με βαθμό αφενός μέχρι 7 και αφετέρου από 7 και πάνω;

Απάντηση:
Έστω Α το ενδεχόμενο επιτυχίας με βαθμό μέχρι 7 και Β το ενδεχόμενο από 7 και πάνω. Οι σχετικές συχνότητές τους, θα χρησιμοποιηθούν ως εκτιμήσεις των αντίστοιχων εμπειρικών πιθανοτήτων, δηλ.:

$P(A)=\frac{n(A)}{N}=\frac{1.975}{5.000}=0,395$ και αντίστοιχα για την $P(B)=\frac{n(B)}{N}=\frac{1.230}{5.000}=0,246$

Σε αντίθεση με τις παραπάνω ερμηνείες της πιθανότητας οι de Finetti, Savage, Lindey κ.ά. υποστηρίζουν ότι δεν υπάρχει αντικειμενική πιθανότητα για ένα ενδεχόμενο. Αντίθετα, θεωρούν ότι υπάρχει μόνο υποκειμενική αξιολόγηση της αβεβαιότητας έκβασης των πειραμάτων τύχης, η οποία προκύπτει από το βαθμό εμπιστοσύνης που οι άνθρωποι αποδίδουν για κάθε ενδεχόμενο του δειγματικού χώρου. Ο τρόπος μέτρησης της υποκειμενικής πιθανότητας αποτυπώνεται από το ποσό στο οποίο μπορεί κάποιος να στοιχηματίσει υπέρ της εμφάνισης του υπό εξέταση ενδεχομένου.

Παράδειγμα 5.8

Μια ομάδα ερευνητών βρίσκεται σε δίλημμα αν πρέπει να συντάξει πρόταση για χρηματοδότηση ερευνητικού της έργου ή να διαθέσει το χρόνο αυτό για να υποβάλλει σχετικά άρθρα προς ανώνυμη-τυφλή κρίση και δημοσίευση, με βάσει το πλήθος και την ποιότητα των οποίων χρηματοδοτείται από το ΥΠΕΠΘ. Ο επιστημονικός της υπεύθυνος στοιχηματίζει, αυτήν τη φορά, υπέρ της αποδοχής της πρότασης για χρηματοδότηση από το υπουργείο με 500 ευρώ έναντι 200 ευρώ στην αντίθετη περίπτωση. Ποια είναι η εκτίμηση της πιθανότητας επιτυχούς κατάληξης της πρότασης για χρηματοδότηση ερευνητικού έργου της υπόψη ομάδας, από τον επιστημονικό της υπεύθυνο;

Απάντηση:

Έστω ότι συμβολίζουμε με Α το ενδεχόμενο επιτυχούς κατάληξης της πρότασης για χρηματοδότηση ερευνητικού έργου και με Β το ενδεχόμενο απόρριψης της πρότασης για χρηματοδότηση από το ΥΠΕΠΘ. Επομένως, ως ξένα τα δύο ενδεχόμενα που καλύπτουν πλήρως το δειγματικό χώρο, θα έχουμε Α = S-B και Β = S-A.

Αφού ο επιστημονικός υπεύθυνος στοιχηματίζει 500 έναντι 200 ευρώ για τα Α και Β αντίστοιχα, τότε ισοδύναμα μπορούμε να γράψουμε:

$$\frac{P(A)}{1-P(A)} = \frac{500}{200} \Rightarrow P(A) = \frac{5}{7} = 0,7143$$

Συμπερασματικά, για τον αριθμό P(A) που ονομάσαμε πιθανότητα του ενδεχομένου Α το οποίο ανήκει (είναι υποσύνολο) του δειγματικού χώρου S πρέπει να θυμόμαστε ΠΑΝΤΑ τα εξής αξιώματα:

$P(A) \geq 0, \forall A$	**(5.6)**

$P(S) = 1$	**(5.7)**

$P(A\ \text{ή}\ \Gamma) = P(A) + P(\Gamma)$	**(5.8)**

Όπου Α, Γ ξένα ενδεχόμενα

$0 \leq P(A) \leq 1$	**(5.9)**

$P(A') = 1 - P(A), \text{αφού}\ S = A \cup A',\ A \cap A' = \varnothing\ \text{και}\ P(\varnothing) = 0 \Rightarrow$ $P(S) = P(A) + P(A') = 1 \Rightarrow P(A') = 1 - P(A)$	**(5.10)**

Όπου η σχέση (5.6) λέει ότι οι πιθανότητες παίρνουν μόνο θετικές τιμές ή μηδέν, ενώ η σχέση (5.9) δίνει το εύρος διακύμανσή τους. Η σχέση (5.6) λέει ότι η πιθανότητα εμφάνισης του δειγματικού χώρου είναι βέβαιο ενδεχόμενο, ενώ οι ενσωματωμένες στην (5.10) αναφέρονται στις σχέσεις που αναφέρονται στο ενδεχόμενο Α και το συμπλήρωμά του Α'.

Παράδειγμα 5.9

Με βάση τα στοιχεία του Πίνακα 5.1 για τη συμπεριφορά δείγματος 500 φοιτητών αναφορικά με τα ΠΜΣ, ζητείται να κατασκευάσετε πίνακα συνάφειας των πιθανοτήτων των απλών και σύνθετων ενδεχομένων από τα δεδομένα του πίνακα 5.1, υπολογίζοντας τις οριακές (marginal probabilities) και συνδυασμένες (join probabilities) πιθανότητες.

Απάντηση:

Ορίζουμε τα απλά ενδεχόμενα:

- A = ΝΑΙ, σχεδιάζω να παρακολουθήσω ΠΜΣ
- A′ = ΟΧΙ, δεν σχεδιάζω να παρακολουθήσω ΠΜΣ
- B = ΝΑΙ, συνεχίζω σε ΠΜΣ
- B′ = ΟΧΙ, δεν συνεχίζω σε ΠΜΣ

Επομένως οι ζητούμενες οριακές (περιθώριες) στήλη και γραμμή του πίνακα συνάφειας, -σύνολα- ή απλές πιθανότητες θα είναι:

- P(A) = 125/500 = 0,250
- P(A′) = 375/500 = 0,750
- P(B) = 150/500 = 0,300
- P(B′) = 350/500 = 0,700

Αναφορικά με τις συνδυασμένες πιθανότητες χρησιμοποιούμε τους ορισμούς της αντικειμενικής πιθανότητας (σχέσεις 5.4 ή 5.5) και έχουμε:

- P(A και B) = 100/500 = 0,200
- P(A και B′) = 25/500 = 0,050
- P(A′ και B) = 50/500 = 0,100
- P(A′ και B′) = 325/500 = 0,650

Τα παραπάνω συνοψίζονται στον Πίνακα 5.2.

Πίνακας 5.2 Πίνακας συνάφειας απλών και συνδυασμένων πιθανοτήτων

	Συνεχίζουν σε ΠΜΣ		
Σχεδιάζουν για ΠΜΣ	ΝΑΙ	ΌΧΙ	Σύνολο
ΝΑΙ	0,200	0,050	0,250
ΌΧΙ	0,100	0,650	0,750
Σύνολο	0,300	0,700	1,000

Από το Παράδειγμα 5.9 πρέπει να έχει γίνει σαφές ότι το άθροισμα των συνδυασμένων πιθανοτήτων ισούται με την περιθώρια, όταν πρόκειται για μετρήσεις αμοιβαία αποκλειόμενων ενδεχομένων που εξαντλούν το δειγματικό χώρο. Για παράδειγμα P(A) = P(A και B) + P(A και B′) = 0,200 + 0,050 = 0,250.

Έτσι, οδηγούμαστε στον ορισμό της απλής ή οριακής πιθανότητας ως άθροισμα συνδυασμένων:

$P(A) = P(A\,\kappa\alpha\iota\,B_1) + P(A\,\kappa\alpha\iota\,B_2) + \ldots + P(A\,\kappa\alpha\iota\,B_N)$	**(5.11)**

Όπου
B_1, B_2, ..., B_N είναι N αμοιβαία αποκλειόμενα ενδεχόμενα (mutually exclusive) που εξαντλούν το δειγματικό χώρο (collectively exhaustive events).

Μία άλλη έννοια πιθανότητας, πολύ σημαντική γιατί είναι ευρέως χρησιμοποιούμενη στη στατιστική επαγωγή, είναι η «πιθανότητα του ενδεχομένου A ή B» P(A ή B). Αυτή η πιθανότητα ορίζεται είτε ως το ενδεχόμενο A είτε ως το ενδεχόμενο B είτε ως και τα δύο ενδεχόμενα A και B. Συμβολικά, μιλάμε για το γενικό προσθετικό κανόνα των πιθανοτήτων και γράφουμε:

$P(A\,\acute{\eta}\,B) = P(A) + P(B) - P(A\,\kappa\alpha\iota\,B)$ ή $P(A \cup B) = P(A) + P(B) - P(A \cap B)$	**(5.12)**

Εντούτοις, αν τα ενδεχόμενα A και B είναι αμοιβαία αποκλειόμενα, επειδή τότε η συνδυασμένη τους πιθανότητα είναι μηδενική, δηλ., ο γενικός προσθετικός κανόνας τροποποιείται στο λεγόμενο ειδικό κανόνα της πρόσθεσης πιθανοτήτων ως εξής:

$P(A\,\acute{\eta}\,B) = P(A) + P(B)$ ή $P(A \cup B) = P(A) + P(B)$	**(5.13)**

Παράδειγμα 5.10

Συνεχίζοντας το προηγούμενο παράδειγμα, με βάση τα στοιχεία του Πίνακα 5.1 για τη συμπεριφορά δείγματος 500 φοιτητών αναφορικά με τα ΠΜΣ, ποια είναι η πιθανότητα να σχεδίαζε κάποιος από αυτούς να παρακολουθήσει ΠΜΣ ή ήδη να φοιτά σε αυτό;

Απάντηση:

Με βάση τον ορισμό που δώσαμε προηγουμένως για τα απλά ενδεχόμενα:

- A = ΝΑΙ, σχεδιάζω να παρακολουθήσω ΠΜΣ
- A′ = ΟΧΙ, δεν σχεδιάζω να παρακολουθήσω ΠΜΣ
- B = ΝΑΙ, συνεχίζω σε ΠΜΣ, σύμφωνα με το σχεδιασμό μου
- B′ = ΟΧΙ, δεν συνεχίζω σε ΠΜΣ, παρά το σχεδιασμό μου

μας ζητείται η P(A ή B) η οποία βάσει του γενικού κανόνα της πρόσθεσης υπολογίζεται ως εξής:

P(A ή B) = P(A) + P(B) - P(A και B) = 0,250 + 0,300 – 0,200 = 0,350.

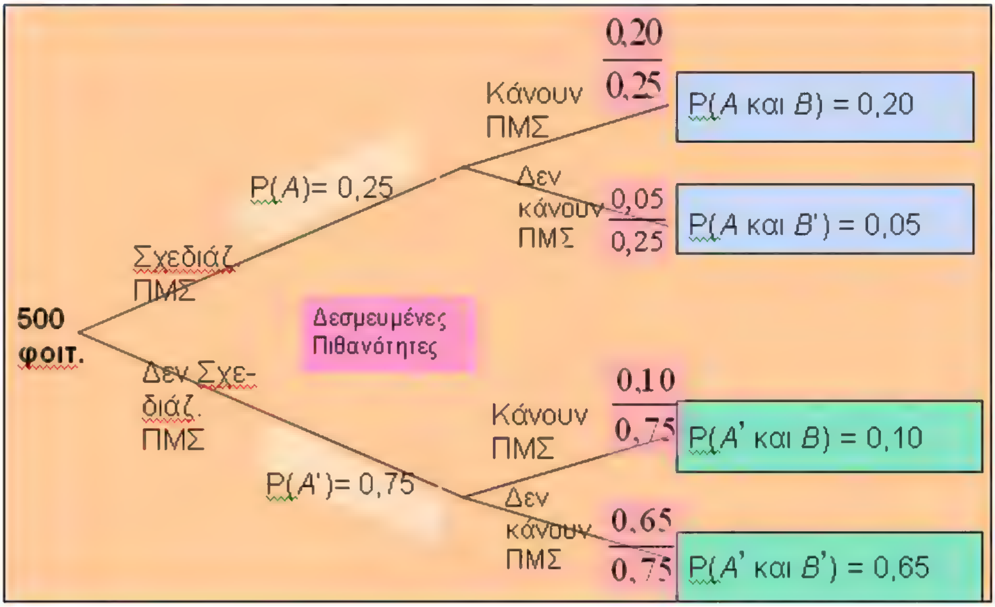
Με άλλα λόγια η πιθανότητα κάποιος φοιτητής να σχεδίαζε να παρακολουθήσει ΠΜΣ ή πράγματι να φοιτά σε κάποιο είναι 35%.

Εάν, αντίθετα με τον τρόπο που μέχρι τώρα δειγματοληπτούμε, δηλ. από ολόκληρο το δειγματικό χώρο, ζητείται να υπολογίσουμε την πιθανότητα εμφάνισης ενδεχομένου, εάν ορισμένη πληροφορία για άλλα ενδεχόμενα του υπόψη δειγματικού χώρου είναι ήδη γνωστή, τότε, καταρχάς, αναφερόμαστε στην έννοια της δεσμευμένης ή υπό συνθήκη (περιορισμό) πιθανότητας του ενδεχομένου A δεδομένης της πληροφόρησης για το B. Με άλλα λόγια εδώ υπολογίζουμε την πιθανότητα του A στο υποσύνολο B του δειγματικού χώρου S, όχι σε ολόκληρο τον τελευταίο. Συμβολικά η δεσμευμένη πιθανότητα δίνεται από τη σχέση:

$$P(A|B) = \frac{P(A\,\kappa\alpha\iota\,B)}{P(B)} \quad \text{ή} \quad P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \textbf{(5.14)}$$

Οι μέχρι τώρα περιγραφείσες έννοιες της απλής ή οριακής, της συνδυασμένης και της υπό συνθήκη ή δεσμευμένης πιθανότητας μπορεί να απεικονιστούν με σαφή τρόπο σε δενδροδιαγράμματα. Για το παράδειγμα της συμπεριφοράς του δείγματος των 500 φοιτητών για τα ΠΜΣ που χρησιμοποιούμε, το αντίστοιχο δενδροδιάγραμμα μπορεί να είναι:

Σχήμα 5.3 Δενδροδιάγραμμα απλών, συνδυασμένων και δεσμευμένων πιθανοτήτων



Παράδειγμα 5.11

Συνεχίζοντας το παράδειγμα, από τα στοιχεία του Πίνακα 5.1 για τη συμπεριφορά δείγματος 500 φοιτητών αναφορικά με τα ΠΜΣ, ποια είναι η πιθανότητα κάποιος φοιτητής να φοιτά σε ΠΜΣ δεδομένου ότι το σχεδίαζε από την αρχή των προπτυχιακών του σπουδών;

Απάντηση:

Με βάση τον ορισμό που έχουμε δώσει για τα απλά ενδεχόμενα, δηλ.:

- Α = ΝΑΙ, σχεδιάζω να παρακολουθήσω ΠΜΣ
- Α′ = ΟΧΙ, δεν σχεδιάζω να παρακολουθήσω ΠΜΣ
- Β = ΝΑΙ, συνεχίζω σε ΠΜΣ
- Β′ = ΟΧΙ, δεν συνεχίζω σε ΠΜΣ

μας ζητείται η $P(B|A)$ η οποία βάσει του ορισμού της δεσμευμένης πιθανότητας υπολογίζεται ως εξής:

$$P(B|A) = \frac{P(A \text{ και } B)}{P(A)} = \frac{0,200}{0,250} = 0,800$$

Δηλαδή η πιθανότητα κάποιος φοιτητής να υλοποίησε τους σχεδιασμούς του για τα ΠΜΣ είναι 80%.

Συμπερασματικά, από τα παραπάνω παραδείγματα είδαμε ότι η απλή ή οριακή πιθανότητα:

- P(ΝΑΙ, συνεχίζω σε ΠΜΣ) = P(Β) = 0,300

ενώ αντίθετα η δεσμευμένη:

- P(ΝΑΙ, συνεχίζω σε ΠΜΣ|ΝΑΙ, σχεδίαζα να παρακολουθήσω ΠΜΣ)= $P(B|A)$= 0,800.

Αυτό το αποτέλεσμα δείχνει ότι η προηγούμενη γνώση ότι ο φοιτητής σχεδίαζε να κάνει ΠΜΣ επηρέασε την πιθανότητα να το πραγματοποιήσει αυτό. Επομένως, το αποτέλεσμα του Β ενδεχομένου εξαρτάται από το αποτέλεσμα του Α ενδεχομένου.

Αντίθετα, όταν η εμφάνιση ενός ενδεχομένου δεν επηρεάζει την εμφάνιση του άλλου, τότε λέμε ότι είναι στατιστικά ανεξάρτητα. Συμβολικά για τη στατιστική ανεξαρτησία δύο ενδεχομένων, π.χ. Α και Β , χρησιμοποιούμε την παρακάτω εξίσωση

$P(A\|B) = P(A)$	**(5.15)**

Όπου:

$P(A|B)$ η δεσμευμένη πιθανότητα του Α δεδομένου του Β και P(A) η περιθώρια πιθανότητα του Α.

Από τον ορισμό της δεσμευμένης πιθανότητας $P(A|B) = \frac{P(A\,\kappa\alpha\iota\,B)}{P(B)}$ μπορούμε να υπολογίσουμε την υπολογίσουμε τηναντίστοιχη δεσμευμένη P(ΑκαιΒ) εάν γνωρίζουμε την οριακή P(B) και την αντίστοιχη δεσμευμένη $P(A|B)$. Λύνοντας λοιπόν την (5.14) ως προς τη συνδυασμένη, έχουμε αυτό που οι στατιστικολόγοι λένε γενικό κανόνα του πολλαπλασιασμού των πιθανοτήτων.

$P(A\,\kappa\alpha\iota\,B) = P(A\|B)P(B)$ ή $P(A\cap B) = P(A\|B)P(B)$	**(5.16)**

Εντούτοις, για ανεξάρτητα ενδεχόμενα η συνδυασμένη πιθανότητά τους προκύπτει από την αντικατάσταση της δεσμευμένης στην (5.16) από το ίσον της στην (5.15). Από αυτούς τους μετασχηματισμούς προκύπτει ο νόμος του πολλαπλασιασμού για ανεξάρτητα ενδεχόμενα:

$P(A\,\kappa\alpha\iota\,B) = P(A)P(B)$ ή $P(A\cap B) = P(A)P(B)$	**(5.17)**

Από το νόμο του πολλαπλασιασμού (5.16) και τον ορισμό της οριακής πιθανότητας (5.11) προκύπτει εναλλακτικός τρόπος υπολογισμού της τελευταίας με απλούς μετασχηματισμούς των σχέσεων, δηλ.:

Από:

- P(A) = P(A και B_1) + P(A και B_2) +........ + P(A και B_N)
- P(A και B) = $P(A|B)\cdot$ P(B)

αβίαστα προκύπτει ο εναλλακτικός ορισμός της οριακής ή περιθώριας πιθανότητας,

$P(A) = P(A\|B_1)P(B_1) + P(A\|B_2)P(B_2) + \ldots + P(A\|B_N)P(B_N)$	**(5.18)**

Όπου:
B_1, B_2, ..., B_N είναι N αμοιβαία αποκλειόμενα ενδεχόμενα που εξαντλούν το δειγματικό χώρο.

Παράδειγμα 5.12

Στο παράδειγμα του Πίνακα 5.1 για τη συμπεριφορά των 500 φοιτητών αναφορικά με τα ΠΜΣ, ποια είναι η πιθανότητα κάποιος φοιτητής να έχει σχεδιάσει να παρακολουθήσει ΠΜΣ εάν η διαθέσιμη πληροφόρηση αφορά μόνο στο «παρόν» και όχι στο «παρελθόν»;

Απάντηση:

Με βάση τον ορισμό που έχουμε δώσει για τα απλά ενδεχόμενα, δηλ.:

- A = ΝΑΙ, σχεδιάζω να παρακολουθήσω ΠΜΣ
- A′ = ΟΧΙ, δεν σχεδιάζω να παρακολουθήσω ΠΜΣ
- B = ΝΑΙ, συνεχίζω σε ΠΜΣ
- B′ = ΟΧΙ, δεν συνεχίζω σε ΠΜΣ

μας ζητείται η P(A) η οποία μπορεί να υπολογιστεί βάσει του εναλλακτικού ορισμού της οριακής [σχέση (5.18)] ως εξής:

$$P(A) = P(A|B)P(B) + P(A|B')P(B')$$
$$= (0,20/0,30)(0,30) + (0,05/0,70)(0,70) = 0,25$$

Δηλαδή η πιθανότητα κάποιος φοιτητής να είχε σχεδιάσει να κάνει μεταπτυχιακές σπουδές P(A) δεδομένου ότι έχουμε σχετική πληροφορία μόνο από σημερινούς φοιτητές ή μη είναι 25%.

5.5 Τυχαίες μεταβλητές και συναρτήσεις πιθανότητας.

Στην πιθανοθεωρητική ανάλυση των πραγματικών φαινομένων της οικονομικής ή κοινωνικής ζωής, βασικά χρησιμοποιούμε τις έννοιες της τυχαίας μεταβλητής και των συναρτήσεων πιθανότητάς τους.

Ο τρόπος αντιστοίχισης πραγματικών αριθμών σε κάθε ενδεχόμενο του δειγματικού χώρου S ονομάζεται τυχαία μεταβλητή. Πιο συγκεκριμένα, η μονοσήμαντη συνάρτηση με πεδίο ορισμού το δειγματικό χώρο S και πεδίο τιμών υποσύνολο των πραγματικών αριθμών ονομάζεται τυχαία μεταβλητή (τ.μ.).

Είναι προφανές ότι παρά το γεγονός ότι μιλάμε για μεταβλητή, στην ουσία πρόκειται για συνάρτηση, την οποία όμως, επειδή ενδιαφερόμαστε κατά βάση για το πεδίο τιμών της, καταχρηστικά την ονομάζουμε μεταβλητή, ενώ επειδή η εμφάνιση των ενδεχομένων είναι αποτέλεσμα του πειράματος τύχης που παράγει τον υπόψη δειγματικό χώρο, ονομάζονται και τυχαίες.

Με κεφαλαία γράμματα, συνήθως, συμβολίζονται οι τυχαίες μεταβλητές, π.χ. X και με μικρά οι τιμές τους, π.χ. x_i, i=1,2,3,…,n.

Οι τυχαίες μεταβλητές, όπως και οι απλές αριθμητικές διακρίνονται σε ασυνεχείς και συνεχείς ανάλογα με το εάν οι τιμές τους είναι πεπερασμένο σύνολο (ή απαριθμήσιμο) και αντίστοιχα όχι. Για παράδειγμα, η τυχαία μεταβλητή που μετράει το πλήθος των τίτλων σπουδών υποψηφίων ΠΜΣ είναι ασυνεχής, ενώ εκείνη που μετράει το οικογενειακό τους εισόδημα είναι συνεχής.

Η συνάρτηση που ως πεδίο ορισμού έχει τις τιμές x_i, i=1,2,3,…,n, π.χ. διακριτής τ.μ. X, και πεδίο τιμών τις αντίστοιχες πιθανότητές τους $f(x_i)$ ονομάζεται συνάρτηση πιθανότητας της ασυνεχούς τ.μ. X και γράφουμε:

$f(x_i) = P_{x_i} = P(x_i) = P(X = x_i), \quad i = 1, 2, \ldots, n$ αν ικανοποιούνται οι συνθήκες $f(x_i) \geq 0 \quad \kappa\alpha\iota \quad \sum_i f(x_i) = 1$	**(5.19)**

Η αντίστοιχη έννοια για τις συνεχείς τ.μ. ονομάζεται συνάρτηση πυκνότητας πιθανότητας και ορίζεται ως εξής:

$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)dx, \text{με } \alpha < \beta$ αν ικανοποιούνται οι συνθήκες $f(x) \geq 0 \quad, -\infty < x < +\infty \quad \kappa\alpha\iota \quad \int_{-\infty}^{+\infty} f(x)dx = 1$	**(5.20)**

Η καταγραφή τιμών με τις αντίστοιχες πιθανότητές τους σε Πίνακα ανάλογης μορφής του Πίνακα συχνοτήτων στην Περιγραφική Στατιστική, ονομάζεται Κατανομή Πιθανότητας της ασυνεχούς τυχαίας μεταβλητής. Επίσης, ανάλογη έννοια με την αντίστοιχη της αθροιστικής συνάρτησης και Πίνακα αθροιστικών συχνοτήτων της Περιγραφικής Στατιστικής είναι η Συνάρτηση Κατανομής ή αθροιστική συνάρτηση κατανομής με τον αντίστοιχο Πίνακα.

$F(x) = P(-\infty < X \leq x) = \begin{cases} \sum_{X \leq x} f(x), & \alpha\nu\ X\ \delta\iota\alpha\kappa\rho\iota\tau\acute{\eta} \\ \int_{-\infty}^{x} f(x)dx, & \alpha\nu\ X\ \sigma\upsilon\nu\varepsilon\chi\acute{\eta}\varsigma \end{cases}$	**(5.21)**

Αν ικανοποιούνται οι συνθήκες:

α) F(x) μη-φθίνουσα του x , δηλ. αν $x_1<x_2$ τότε και $F(x_1) \leq F(x_2)$.

β) min F(x)=0 και max F(x)=1, δηλ. $F(-\infty)=0$ και $F(+\infty)=1$.

γ) F(x) συνεχής από τα δεξιά για όλα τα $x \in \Re$

Παράδειγμα 5.13

Έστω Χ ορίζουμε την τυχαία μεταβλητή (τ.μ.) «αριθμός Κοριτσιών σε 4 συνεχόμενες γεννήσεις ορισμένης μαιευτικής κλινικής». Να δείξετε α) πώς υπολογίζονται οι τιμές και οι πιθανότητές της, καθώς επίσης και πως κατασκευάζεται η κατανομή πιθανότητας της, τόσο η απλή όσο και η αθροιστική και β) F(3)=; γ) P(X>1)=; δ) P(1≤X≤3)=;

Απάντηση:

α) Το ενδεχόμενο της γέννησης στην πρώτη γέννηση έστω αγοριού δεν επηρεάζει το αποτέλεσμα της δεύτερης. Επομένως, τα ενδεχόμενα των γεννήσεων είναι ανεξάρτητα. Δεδομένου ότι τα δυνατά ενδεχόμενα σε κάθε γέννηση είναι m=2, δηλ. αγόρι (Α) ή κορίτσι (Κ), το σύνολο του δειγματικού χώρου στις τέσσερις γεννήσεις («επαναλήψεις του πειράματος τύχης=n») θα είναι $m_n=2^4=16$. Οι διατάξεις των ενδεχομένων αυτών, οι απαριθμήσεις των κοριτσιών ανά ενδεχόμενο, δηλ. οι τιμές (x_i), της τυχαίας μας μεταβλητής έστω Χ, μαζί με τις αντίστοιχες πιθανότητές τους, (βάσει του νόμου του πολλαπλασιασμού ανεξάρτητων ενδεχομένων), δίνονται στον παρακάτω Πίνακα

Πίνακας 5.3 Υπολογισμός τιμών και πιθανοτήτων τυχαίας μεταβλητής.

	Στοιχειώδη ενδεχόμε-να	Τιμές τ.μ. (x_i)	Πιθανότητες $f(x)=P(X=x)$
1	ΑΑΑΑ	0	P(A∩A∩A∩A) = P(A) P(A) P(A) P(A) = (½)(½)(½)(½) = 1/16
2	ΑΑΑΚ	1	P(A∩A∩A∩K) = P(A) P(A) P(A) P(K) = (½)(½)(½)(½) = 1/16
3	ΑΑΚΑ	1	P(A∩A∩K∩A) = P(A) P(A) P(K) P(A) = (½)(½)(½)(½) = 1/16
4	ΑΑΚΚ	2	P(A∩A∩K∩K) = P(A) P(A) P(K) P(K) = (½)(½)(½)(½) = 1/16
5	ΑΚΑΑ	1	P(A∩K∩A∩A) = P(A) P(K) P(A) P(A) = (½)(½)(½)(½) = 1/16
6	ΑΚΑΚ	2	P(A∩K∩A∩K) = P(A) P(K) P(A) P(K) = (½)(½)(½)(½) = 1/16
7	ΑΚΚΑ	2	P(A∩K∩K∩A) = P(A) P(K) P(K) P(A) = (½)(½)(½)(½) = 1/16
8	ΑΚΚΚ	3	P(A∩K∩K∩K) = P(A) P(K) P(K) P(K) = (½)(½)(½)(½) = 1/16
9	ΚΑΑΑ	1	P(K∩A∩A∩A) = P(K) P(A) P(A) P(A) = (½)(½)(½)(½) = 1/16
10	ΚΑΑΚ	2	P(K∩A∩A∩K) = P(K) P(A) P(A) P(K) = (½)(½)(½)(½) = 1/16
11	ΚΑΚΑ	2	P(K∩A∩K∩A) = P(K) P(A) P(K) P(A) = (½)(½)(½)(½) = 1/16
12	ΚΑΚΚ	3	P(K∩A∩K∩K) = P(K) P(A) P(K) P(K) = (½)(½)(½)(½) = 1/16
13	ΚΚΑΑ	2	P(K∩K∩A∩A) = P(K) P(K) P(A) P(A) = (½)(½)(½)(½) = 1/16
14	ΚΚΑΚ	3	P(K∩K∩A∩K) = P(K) P(K) P(A) P(K) = (½)(½)(½)(½) = 1/16
15	ΚΚΚΑ	3	P(K∩K∩K∩A) = P(K) P(K) P(K) P(A) = (½)(½)(½)(½) = 1/16
16	ΚΚΚΚ	4	P(K∩K∩K∩K) = P(K) P(K) P(K) P(K) = (½)(½)(½)(½) = 1/16

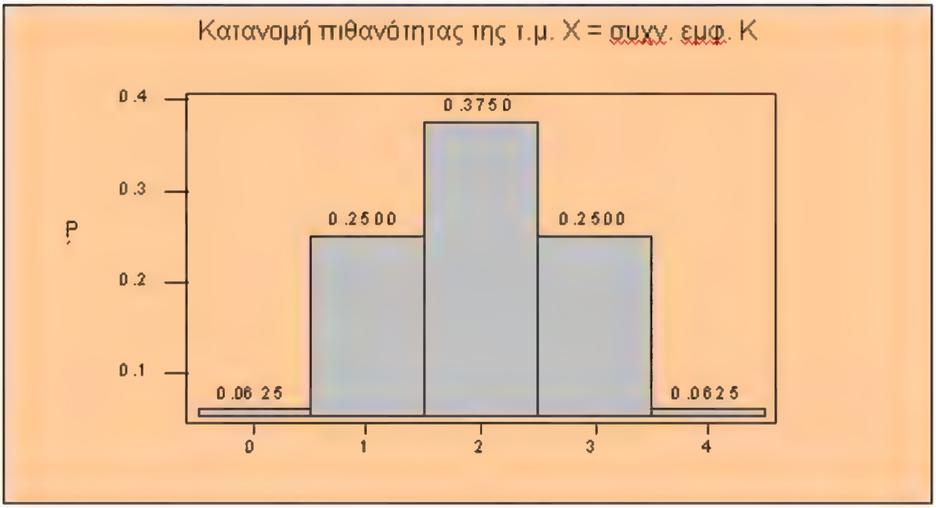
Όπως φαίνεται στον παραπάνω αναλυτικό πίνακα οι τιμές που είναι δυνατόν να πάρει η τυχαία μας μεταβλητή Χ είναι x_i={0, 1, 2, 3, 4}, ενώ οι πιθανότητες κάθε μιας από αυτές προκύπτουν ως άθροισμα των αντίστοιχων στοιχειωδών ενδεχομένων. Για παράδειγμα f(1)=P(X=1)=4/16=1/4 αφού το 1 εμφανίζεται 4 φορές. Τη συνάρτηση πιθανότητας της ασυνεχούς τυχαίας μας μεταβλητής Χ σε μορφή Πίνακα δίνουμε παρακάτω στη λεγόμενη κατανομή πιθανότητας, ανάλογη έννοια της κατανομής συχνότητας στην περιγραφική στατιστική (όπου εκεί έχουμε απόλυτη συχνότητα, εδώ απλές πιθανότητες).

Πίνακας 5.4 Συνάρτηση Πυκνότητας Πιθανότητας και Συνάρτηση Κατανομής της ασυνεχούς τυχαίας μεταβλητής Χ.

Τιμές τ.μ. (x_i)	Πιθανότητες ή Συνάρτηση Πυκν. Πιθαν. $f(x)=P(X=x)$		Τιμές τ.μ. (x_i)	Συνάρτηση Κατανομής $F(x)=P(X\le x)=\sum_{X\le x} f(x)$
0	$f(0)=P(X=0)=1/16$		Μέχρι 0	$F(0)=P(X\le 0)=1/16$
1	$f(1)=P(X=1)=4/16$		Μέχρι 1	$F(1)=P(X\le 1)=5/16$
2	$f(2)=P(X=2)=6/16$		Μέχρι 2	$F(2)=P(X\le 2)=11/16$
3	$f(3)=P(X=3)=4/16$		Μέχρι 3	$F(3)=P(X\le 3)=15/16$
4	$f(4)=P(X=4)=1/16$		Μέχρι 4	$F(4)=P(X\le 4)=16/16$
Σύνολο	$\sum f(x)=16/16=1$			

Η συνάρτηση ή κατανομή πιθανότητας της ασυνεχούς μας τυχαίας μεταβλητής Χ, η οποία παριστά τον αριθμό Κοριτσιών σε 4 συνεχόμενες γεννήσεις ορισμένης μαιευτικής κλινικής, δίνεται διαγραμματικά στο παρακάτω Σχήμα.

Σχήμα 5.4 Ιστόγραμμα Κατανομής Πιθανότητας ασυνεχούς τυχαίας μεταβλητής



β) Με βάση τα στοιχεία των κατανομών πιθανοτήτων (απλής και αθροιστικής) του Πίνακα 5.4 μπορούμε εύκολα να απαντήσουμε στα υπόλοιπα ερωτήματα. Έτσι:

$F(3)=P(X\leq 3)=15/16 =P(X=0)+P(X=1)+P(X=2)+P(X=3)=(1+4+6+4)/16$.

γ) $P(X>1)=F(4)-F(1)=(16/16)-(5/16)=11/16$

δ) $P(1\leq X\leq 3)=F(3)-F(0)= (15/16)-(1/16)=14/16$.

5.6 Βασικές παράμετροι τυχαίων μεταβλητών

Οι παράμετροι ή ροπές (moments) των κατανομών πιθανότητας των τυχαίων μεταβλητών προσδιορίζουν τη μορφή των τελευταίων, όπως οι βασικές στατιστικές του δείγματος τη μορφή της κατανομής ή συνάρτησής του που είδαμε στην περιγραφική στατιστική (τάσης, θέσης, διασποράς, ασυμμετρίας και κύρτωσης).

Οι βασικές ροπές των συναρτήσεων πιθανότητας είναι ο μέσος και η διακύμανση. Ο μέσος είναι η πρώτη ροπή περί την αρχή μηδέν και η διακύμανση η δεύτερη ροπή περί τον μέσο. Οι έννοιες εδώ αναλύονται σε όρους αναμενόμενων τιμών (expected values) ή μαθηματικών ελπίδων (mathematical expectations).

Έτσι, ο μέσος μιας τυχαίας μεταβλητής X με συνάρτηση πυκνότητας πιθανότητας f(x) είναι η αναμενόμενη τιμή της (expected value) E(X), η οποία ορίζεται ως ο μέσος σταθμικός των τιμών της X με στάθμιση τις αντίστοιχες πιθανότητές τους.

$$E(X)=\begin{cases}\sum_i x_i f(x_i)=\mu, & \text{αν X διακριτή}\\ \int_{-\infty}^{+\infty} x f(x)dx=\mu, & \text{αν X συνεχής}\end{cases} \qquad \textbf{(5.22)}$$

Επίσης, αν X τυχαία μεταβλητή με μέσο E(X)=μ, τότε η προσδοκώμενη τιμή των τετραγωνικών αποκλίσεών της από το μέσο, $[X-E(X)]^2$, ονομάζεται διακύμανση της X και ορίζεται ως εξής:

$$\sigma^2=\text{var}(X)=E[X-E(X)]^2=\begin{cases}\sum_i (x_i-\mu)^2 f(x_i)=\mu, & \text{αν X διακριτή}\\ \int_{-\infty}^{+\infty} (x_i-\mu)^2 f(x)dx=\mu, & \text{αν X συνεχής}\end{cases} \qquad \textbf{(5.23)}$$

Για κάθε τ.μ. X μπορεί να δειχτεί ότι ισχύει η παρακάτω ισότητα που διευκολύνει στους υπολογισμούς χωρίς Η/Υ: $var(X) = \sigma^2 = E[X - E(X)]^2 = E(X^2) - [E(X)^2]$.

Βασικές ιδιότητες (θεωρήματα) για το μέσο και τη διακύμανση τυχαίων μεταβλητών δίνονται παρακάτω[5]:

E(α)=α, όπου α=σταθερά	**(5.24)**
E(αX)=αE(X), όπου α=σταθερά και X=τ.μ.	**(5.25)**
E(αX+β)=αE(X)+β, όπου α, β=σταθερές και X=τ.μ.	**(5.26)**
Var(α)=0, όπου α =σταθερά	**(5.27)**
Var(αX+β)=$α^2$var(X), όπου α, β=σταθερές και X=τ.μ.	**(5.28)**

Παράδειγμα 5.14

Για το προηγούμενο παράδειγμα της διακριτής τυχαίας μεταβλητής X, η οποία παριστά τον αριθμό κοριτσιών σε 4 συνεχόμενες γεννήσεις ορισμένης μαιευτικής κλινικής, να εκτιμήσετε το μέσο και τη διακύμανσή της, δίνοντας και τις σχετικές ερμηνείες.

Απάντηση:

Με βάση τους ορισμούς (5.22) και (5.23) οι υπολογισμοί φαίνονται στον παρακάτω πίνακα:

Πίνακας 5.5 Υπολογισμοί για το μέσο και τη διακύμανση της ασυνεχούς τυχαίας μεταβλητής X

Τιμές τ.μ. (x_i)	Πιθανότητες *f(x)=P(X=x)*	xf(x)	$[X-E(X)]^2f(x)$
0	0,0625	0x0,0625=0,000	$(0-2,000)^2$x0,0625=0,250
1	0,2500	1x0,2500=0,250	$(1-2,000)^2$x0,2500=0,250
2	0,3750	2x0,3750=0,750	$(2-2,000)^2$x0,3750=0,000
3	0,2500	3x0,2500=0,750	$(3-2,000)^2$x0,2500=0,250
4	0,0625	4x0,0625=0,250	$(4-2,000)^2$x0,0625=0,250
Άθροισμα	**1,0000**	μ=E(X)=**2,000**	**1,000**

[5] Οι αποδείξεις αφήνονται ως άσκηση στον αναγνώστη. Υπάρχουν φυσικά σε όλα τα γνωστά πανεπιστημιακά εγχειρίδια, πχ. Δ. Λαμπράκης (1980), Στατιστική, αυτό-έκδοση (βλ. βιβλιοπωλείο

ΣΜΠΙΛΙΑΣ-ΤΟ ΟΙΚΟΝΟΜΙΚΟ), κ.λπ.
Επομένως, ο αναμενόμενος αριθμός γεννήσεων κοριτσιών στις 4 συνεχόμενες είναι E(X)=2 κορίτσια, ή κάτω από τις προϋποθέσεις που θα εξηγήσουμε στα επόμενα, αν θέλαμε με ένα μόνο αντιπροσωπευτικό αριθμό να εκφράσουμε τα 16 ενδεχόμενα του δειγματικού χώρου της X, τότε αυτός θα ήταν ο μέσος 2. Στο παράδειγμα αυτό ο μέσος συμπίπτει και με τον τύπο ή επικρατούσα τιμή, δηλ. ο επικρατέστερος αριθμός ή συνηθέστερα εμφανιζόμενες γεννήσεις κοριτσιών είναι 2 στις 4 συνεχόμενες γεννήσεις.

Επίσης, η διακύμανση της X είναι σ^2=var(X)=1 επομένως και η τυπική απόκλιση (θετική τετραγωνική ρίζα της) σ=1 που σημαίνει ότι η κατά μέσο όρο απόκλιση των γεννήσεων κοριτσιών ανά 4 συνεχόμενες από τη μέση τιμή τους είναι 1. Σε όρους κανονικής κατανομής πιθανότητας (αν και συνεχής) εδώ θα λέγαμε ότι π.χ. υπάρχει πιθανότητα 68% περίπου οι γεννήσεις κοριτσιών να είναι στο διάστημα [μ-1σ, μ+1σ]=[1, 3] ή το 68% των γεννήσεων θα είναι κορίτσια μεταξύ 1 και 3.

Μια άλλη πολύ σημαντική παράμετρος για τη μέτρηση της συμμεταβολής δύο τυχαίων μεταβλητών είναι η συνδιακύμανσή τους. Η συνδιακύμανση π.χ. των διακριτών τυχαίων μεταβλητών X και Y μετράει, αφενός, τη μορφή (θετική ή αρνητική) και αφετέρου, το βαθμό ή την ένταση της γραμμικής συσχέτισης (συνάφειας) τους στις απόλυτες μονάδες μέτρησής τους και συμβολίζεται ως:

$Cov(X,Y) = E\{[x_i - E(X)][y_i - E(Y)]\} = E[(X-\mu_x)(Y-\mu_y)]$ $= \sum_i \sum_j [x_i - E(X)][y_j - E(Y)] f(x_i y_j)$ $= \sum_i \sum_j x_i y_j f(x_i y_j) - \mu_x \mu_y$	**(5.29)**

Όπου

$$\mu_x = \sum_i x_i f(x_i) \ \text{και}\ \mu_y = \sum_j y_j f(y_j)$$

Έτσι, θετική συνδιακύμανση δείχνει θετική συσχέτιση των X, Y, δηλ. όταν αυξάνει η μια,f τότε αυξάνει και η άλλη και αντίστροφα στη μείωση, ενώ αρνητική συνδιακύμανση σημαίνει ότι όταν αυξάνει η X τότε η Y μειώνεται και αντίστροφα. Εάν οι δύο διακριτές μεταβλητές έχουν συνδιακύμανση μηδέν τότε λέμε ότι είναι γραμμικά ασυσχέτιστες.

Ας σημειωθεί όμως ότι η μηδενική συνδιακύμανση είναι αναγκαία και όχι ικανή συνθήκη ανεξαρτησίας δύο τυχαίων μεταβλητών.

Όταν θέλουμε να μετρήσουμε τη συμμεταβολή των X, Y χρησιμοποιώντας στατιστικό απαλλαγμένο των μονάδων μέτρησής τους, δηλ. όταν ψάχνουμε τη συνδιακύμανση σε μονάδες τυπικών τους αποκλίσεων, χρησιμοποιούμε το συντελεστή γραμμικής συσχέτισης (ρ_{XY}) του Pearson που υπολογίζεται από τη σχέση:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_i \sum_j [x_i - E(X)][y_j - E(Y)] f(x_i y_j)}{\sqrt{\sum_{i=1}^{n} [x_i - E(X)] f(x_i)} \sqrt{\sum_{j=1}^{k} [y_j - E(Y)] f(y_j)}} \qquad (5.30)$$

Το εύρος διακύμανσης του συντελεστή συσχέτισης βρίσκεται στο κλειστό διάστημα [-1,+1], δηλ. $-1 \leq \rho_{XY} \leq +1$, μεταφράζοντας από την τέλεια αρνητική (ρ_{XY} =-1) έως την τέλεια θετική (ρ_{XY} =+1) γραμμική συσχέτιση των Χ και Υ, ενώ όταν (ρ_{XY} =0) μιλάμε για ασυσχέτιστες.

Παράδειγμα 5.15

Επενδυτής θέλει να αποφασίσει πού θα επενδύσει 300.000€ για την επόμενη χρονιά. Εξετάζει για το λόγο αυτό τις αποδόσεις 2 εταιρειών παροχής επενδυτικών υπηρεσιών (ΕΠΕΥ) της τελευταίας πενταετίας κάτω από 3 εναλλακτικά σενάρια του οικονομικού κύκλου, ύφεση, στασιμότητα και μεγέθυνση. Οι σχετικές αποδόσεις και οι υποκειμενικές του πιθανότητες δίνονται στον παρακάτω πίνακα. Με βάση τις έννοιες των μέσου, τυπικής απόκλισης συνδιακύμανσης και συντελεστή συσχέτισης, ποια θα μπορούσε να είναι η επενδυτική σας συμβουλή;

Πίνακας 5.6 Αποδόσεις και υποκειμενικές πιθανότητες επενδυτή

		Αποδόσεις € (ανά 100.000€)	
$P(x_i y_i)$	Φάση οικον. κύκλου	ΕΠΕΥ_Ι (Χ)	ΕΠΕΥ_ΙΙ (Υ)
0,30	ύφεση	-10.000	5.000
0,50	στασιμότητα	10.000	8.000
0,20	μεγέθυνση	35.000	2.000

Απάντηση:

Ορίζουμε X την απόδοση της ΕΠΕΥ_Ι και Υ την απόδοση της ΕΠΕΥ_ΙΙ. Εφαρμόζοντας τις παραπάνω σχέσεις, ενώ οι υπολογισμοί φαίνονται στον Πίνακα 5.7, έχουμε:

Πίνακας 5.7 Υπολογισμοί για την εκτίμηση E(X), E(Y), σ_X, σ_Y, Cov(X,Y), ρ_{XY}

		Αποδόσεις € (ανά 100.000€)						
$P(x_iy_i)$	Φάση οικον. κύκλου	ΕΠΕΥ_Ι (X)	ΕΠΕΥ_ΙΙ (Y)	xP(x)	yP(y)	$[x\text{-}E(X)]^2$ P(x)	$[y\text{-}E(Y)]^2$ P(y)	[x-E(X)] [y-E(Y)]P(xy)
0,30	ύφεση	-10.000	5.000	-3.000	1.500	108.300.000	243.000	5.130.000
0,50	σταθερότητα	10.000	8.000	5.000	4.000	500.000	2.205.000	1.050.000
0,20	μεγέθυνση	35.000	2.000	7.000	400	135.200.000	3.042.000	-20.280.000
	Σύνολα			**9.000**	**5.900**	**244.000.000**	**5.490.000**	**-14.100.000**

E(X)=9.000€, E(Y)=5.900€
Από εδώ εκτιμάμε ότι η κατά μέσο όρο απόδοση από τη διαχείριση που κάνει η ΕΠΕΥ_Ι είναι πολύ υψηλότερη (+53% περίπου) της άλλης (ΕΠΕΥ_ΙΙ).

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{244.000.000} = 15.620$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{5.490.000} = 2.343$$

Αυτοί οι υπολογισμοί μας οδηγούν στην εκτίμηση ότι ο κίνδυνος (που εδώ αξιολογείται από την τυπική απόκλιση των αποδόσεων) από τη διαχείριση που κάνει η ΕΠΕΥ_Ι είναι πολλαπλάσιος (6,7 φορές περίπου) εκείνου της ΕΠΕΥ_ΙΙ.

Cov(X,Y)=-14.100.000

Από τη συνδιακύμανση βλέπουμε ότι οι αποδόσεις που πετυχαίνουν (ιστορικά) οι δύο ΕΠΕΥ έχουν αρνητική συσχέτιση ή ότι κυμαίνονται σε αντίθετες κατευθύνσεις.

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = -0,385$$

Από το συντελεστή συσχέτισης των αποδόσεων που πέτυχαν οι δύο ΕΠΕΥ την περασμένη πενταετία παρατηρούμε μέτρια αρνητική συσχέτιση, δηλ. σε μια κατά 1% αύξηση της απόδοσης της πρώτης η άλλη παρουσίαζε πτώση κατά 0,385%.

Συμπερασματικά, θα λέγαμε ότι για συντηρητικούς επενδυτές η ΕΠΕΥ_ΙΙ είναι ελκυστικότερη αφού παρουσιάζει μικρότερες διακυμάνσεις αποδόσεων και επομένως μικρότερο κίνδυνο. Αντίθετα, για τους ριψοκίνδυνους θα συνιστούσαμε την ΕΠΕΥ_Ι με πολύ υψηλότερη (ιστορικά) απόδοση αλλά και πολλαπλάσιο κίνδυνο. Εναλλακτικά τρίτη κατηγορία επενδυτών θα μπορούσε να επενδύσει τμήμα του κεφαλαίου της στην Ι και το υπόλοιπο στη ΙΙ.

6. Εκπαιδευτική Ενότητα

- **Στοιχεία Πιθανοθεωρίας ΙΙ**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατανοεί τα θεωρητικά Στατιστικά Πρότυπα.
- Να υπολογίζει διωνυμικές πιθανότητες.
- Να εφαρμόζει τις παραμέτρους και να κατανοεί τη μορφή της Διωνυμικής κατανομής.
- Να κατανοεί την έννοια της κανονικής κατανομής και να χρησιμοποιεί τις ιδιότητές της.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Διωνυμική κατανομή.
- Διωνυμικός νόμος πιθανοτήτων
- Κανονική κατανομή
- Κρίσιμες κανονικών τυχαίων μεταβλητών

6.1 Εισαγωγή

Εξηγήσαμε ήδη στο προηγούμενο κεφάλαιο τη σπουδαιότητα των τυχαίων μεταβλητών στην ανάλυση των στοχαστικών φαινομένων του πολύπλοκου κόσμου μας.

Εντούτοις, η συνάρτηση πυκνότητας πιθανότητας και η συνάρτηση κατανομής που περιγράφουν πλήρως τις τυχαίες μεταβλητές (που μέχρι κάποια τιμή[6] τους αντιστοιχίζουν την πιθανότητα εμφάνισής της), δεν είναι δυνατόν να υπολογίζονται, ως αλγεβρική μορφή, για κάθε περίπτωση ξεχωριστά. Άλλωστε, είναι συνήθως σπουδαιότερο να γνωρίζουμε το γενικό πιθανοθεωρητικό πρότυπο συμπεριφοράς κάποιου φαινομένου, παρά την πιθανότητα που παίρνει κάποια τιμή ή διάστημα τιμών ορισμένης τυχαίας μεταβλητής.

Για τους λόγους αυτούς, οι στατιστικολόγοι έχουν διατυπώσει πρότυπα συναρτήσεων πυκνότητας πιθανότητας που έχουν τόσο μεγάλη εφαρμογή σε πλήθος στοχαστικών φαινομένων, έτσι ώστε αφενός, έχουν πινακοποιηθεί και είναι διαθέσιμοι στα εγχειρίδια στατιστικής ή τα σχετικά λογισμικά (ακόμα και στο MS-Excel), και αφετέρου, αναφέρονται στη βιβλιογραφία ως θεωρητικές κατανομές τυχαίων μεταβλητών.

Η **στατιστική συμπερασματολογία** (statistical inference) δηλ., η γενίκευση των συμπερασμάτων του τυχαίου και αντιπροσωπευτικού δείγματος, στο γεννήτορα πληθυσμό του, διευκολύνεται ιδιαίτερα με τη γνώση της συμπεριφοράς των τυχαίων μεταβλητών, πολλές από τις οποίες περιγράφονται αναλυτικά από τους υπόψη θεωρητικούς νόμους πιθανότητας.

Επίσης, είναι ήδη γνωστό ότι οι τυχαίες μεταβλητές διακρίνονται σε ασυνεχείς ή διακριτές και συνεχείς. Κατ' επέκταση, και οι θεωρητικές συναρτήσεις πυκνότητας πιθανότητας ή οι συναρτήσεις κατανομής διακρίνονται και αυτές σε ασυνεχείς και συνεχείς.

Στο κεφάλαιο αυτό θα παρουσιάσουμε σύντομα μόνο δύο, αλλά ακρογωνιαίους λίθους, από τους γνωστούς θεωρητικούς νόμους πιθανότητας, δηλ. την ασυνεχή διωνυμική (Τμήμα 6.2), και τη συνεχή κανονική (Τμήμα 6.3), κατανομές πιθανότητας.

6.2 Διωνυμική κατανομή

6.2.1 Η έννοια του διωνυμικού νόμου πιθανότητας

Ορίζουμε ως Πείραμα Bernoulli το τυχαίο πείραμα με τα χαρακτηριστικά:

- ο δειγματικός χώρος (S) εξαντλείται από δύο μόνο, αμοιβαία αποκλειόμενα ενδεχόμενα τα οποία συμβατικά ονομάζονται επιτυχία[7] (success) και αποτυχία (failure).

- Έτσι εάν Α, τυχαία μεταβλητή που συμβολίζει την επιτυχία και αντίστοιχα Β εκείνη της αποτυχίας, τότε $A \cup B = S$.

- Η πιθανότητα επιτυχίας παραμένει σταθερή σε όλες τις επαναλήψεις του τυχαίου πειράματος, δηλ. P(A)=p, p=σταθερά.

- Οι n επαναλήψεις του πειράματος Bernoulli είναι μεταξύ τους ανεξάρτητες.

---

[6] Εννοούμε εδώ ασυνεχούς τυχαίας.
[7] Η οποία δεν είναι υποχρεωτικό να είναι και το «επιθυμητό αποτέλεσμα».

Η διακριτή τυχαία μεταβλητή που μετρά το πλήθος των επιτυχιών σε n ανεξάρτητες επαναλήψεις, ονομάζεται διωνυμική (binomial), ενώ, η συνάρτηση πιθανότητάς της μπορεί να παρουσιαστεί από τη διωνυμική κατανομή πιθανότητας.

Πιο συγκεκριμένα θα λέμε ότι η διακριτή τυχαία μεταβλητή X ακολουθεί τη διωνυμική κατανομή με παραμέτρους n επαναλήψεις και p σταθερή πιθανότητα «επιτυχίας», συμβολίζοντας X~B(n,p) όταν έχει συνάρτηση πιθανότητας:

$f(x) = P(x;n,p) = \binom{n}{x} p^x q^{n-x}$ ή $p_x = \frac{n!}{x!(n-x)!} p^x q^{n-x}$	**(6.1)**

Όπου x=0,1,2,..., n αριθμός επιτυχιών σε n ανεξάρτητες επαναλήψεις του πειράματος και p+q=1, ενώ q παριστά τον αριθμό των «αποτυχιών». Προφανώς ισχύει ότι $0 \le p \le 1$,

ενώ $\binom{n}{x}$ συμβολίζει τους συνδυασμούς n στοιχείων ανά x . Εννοείται ότι πρέπει να ικανοποιούνται[8] οι συνθήκες (5.19), δηλ. $f(x_i) \ge 0$ και $\sum_i f(x_i) = 1$ για i=1, 2,.....n για να είναι η (6.1) συνάρτηση πυκνότητας πιθανότητας.

Πρέπει να είναι σαφές ότι η διωνυμική ικανοποιεί όλα τα χαρακτηριστικά του πειράματος Bernoulli και έχει πάρα πολλές εφαρμογές σε διχοτομημένους πληθυσμούς. Παραδείγματα διωνυμικών μεταβλητών, που τα ενδεχόμενα τους εξαντλούν το δειγματικό χώρο, μπορεί να είναι: «μηδέν-ένα», «επιτυχία-αποτυχία», «κέρδη-ζημιές», «παρόντες-απόντες» (φοιτητές κατά την παρακολούθηση των μαθημάτων), «ανάκαμψη-ύφεση» (της οικονομίας), «ναι-όχι» (αν θα αγοράσει το προϊόν ερωτηθείς καταναλωτής σε έρευνα Marketing), «αγόρι-κορίτσι», «αποδεκτό-ελαττωματικό» (προϊόν εταιρείας σε στατιστικό έλεγχο ποιότητας), «υποτίμηση-ανατίμηση» (εθνικού νομίσματος και επίδρασή του στο εμπορικό ισοζύγιο), διεκδικώ «να πάρω ή όχι» κάποιο συμβόλαιο έργου, κ.λπ.

[8] Όπως έχουμε αναφέρει αλλού, οι αποδείξεις αφήνονται ως άσκηση στον αναγνώστη, ο οποίος μπορεί να συμβουλευτεί όλα τα γνωστά εγχειρίδια επαγωγικής στατιστικής, πχ. Δ.Τερζάκης (1999), Στατιστική Επιχειρήσεων, Interbooks, ή Ε.Θαλασσινός, Θ.Β.Σταματόπουλος, και Χ.Φ.Χαρίσης (1996), Επιχει-ρησιακή Στατιστική, Σταμούλης, ή Α.Κιντής (1994), Στατιστικές και Οικονομετρικές Μέθοδοι, Guten-berg

6.2.2 Υπολογισμός διωνυμικών πιθανοτήτων

Η σχέση (6.1) δίνει τις μεμονωμένες πιθανότητες x «επιτυχιών» σε n επαναλήψεις του πειράματος. Εντούτοις, για μικρό αριθμό επιτυχιών για τις οποίες επιθυμούμε να εκτιμήσουμε τις αντίστοιχες πιθανότητες εμφάνισής τους χρησιμοποιείται εναλλακτικά ο παρακάτω αναδρομικός τύπος.

$$p_{x+1} = \frac{n-x}{x+1}\frac{p}{q}p_x \qquad (6.2)$$

Ήδη από την (6.1) αβίαστα προκύπτει ότι για x=0, $p_0=q^n$. Έτσι με το δεδομένο αυτό και την (6.2) υπολογίζουμε τις πιθανότητες για οποιοδήποτε αριθμό επιτυχιών x επιθυμούμε.

Παράδειγμα 6.1

Κάλπη περιέχει 8 σφαιρίδια, 5 κόκκινα και 3 κίτρινα. Η δειγματοληψία γίνεται με επανάθεση, δηλ. κάθε φορά που επιλέγεται τυχαία 1 σφαιρίδιο στη συνέχεια, πριν την επιλογή του επόμενου το πρώτο επανατοποθετείται στην κάλπη. Αν το πείραμα επιλογής σφαιριδίων γίνει 5 φορές και όπου X συμβολίζουμε την επιλογή κόκκινου σφαιριδίου ζητείται να υπολογιστούν:

α) οι ατομικές πιθανότητες για όλους τους δυνατούς αριθμούς x «επιτυχιών»,

β) οι αθροιστικές πιθανότητες $P(X\leq4)$, $P(X>2)$ και $P(2\leq X\leq4)$.

Απάντηση:

Προφανώς ικανοποιούνται οι υποθέσεις του πειράματος Bernoulli και η X~B(n=5,p=5/8=0,625), q=3/8=0,375 επομένως p + q = 0,625 + 0,375 = 1,000 και p/q=0,625/0,375=1,6667.

Χρησιμοποιώντας τον αναδρομικό τύπο (6.2) θα έχουμε:

$$p_0 = q^n = 0,375^5 = 0,0074$$

$$p_{x+1} = \frac{n-x}{x+1} \cdot \frac{p}{q} \cdot p_x \Leftrightarrow p_{0+1} = \frac{5-0}{0+1} \cdot 1,6667 \cdot 0,0074 \Rightarrow p_1 = 0,0618$$

$$p_2 = \frac{5-1}{1+1} \cdot 1,6667 \cdot 0,0618 \Rightarrow p_2 = 0,2060$$

$$p_3 = \frac{5-2}{2+1} \cdot 1,6667 \cdot 0,2060 \Rightarrow p_3 = 0,3433$$

$$p_4 = \frac{5-3}{3+1} \cdot 1,6667 \cdot 0,3433 \Rightarrow p_4 = 0,2861$$

$$p_5 = \frac{5-4}{4+1} \cdot 1,6667 \cdot 0,2861 \Rightarrow p_5 = 0,0954$$

Εναλλακτικά απευθείας από τη συνάρτηση πιθανότητας (6.1) π.χ. για x=3 θα είχαμε:

$$p_x = \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x} = \frac{5!}{3!(5-3)!} \cdot 0,625^3 \cdot 0,375^{5-3} = 0,3433$$

Εναλλακτικά μπορούσαμε να χρησιμοποιήσουμε τους Πίνακες της Διωνυμικής για n=5, p=0,625 και σταδιακά x=0,1,…,5.

Αυτούς τους Πίνακες της Διωνυμικής κατανομής πιθανότητας χρησιμοποιεί και το MS-Excel στη συνάρτηση «=BINOMDIST(Number_s;trials;prob.;Cumulative Prob. Function=FALSE)», όπου, Number_s είναι το πλήθος των ζητούμενων επιτυχιών x (για το παράδειγμά μας x=0,1,…,5), trials είναι οι n επαναλήψεις, prob. είναι η πιθανότητα επιτυχίας p και το τελευταίο πεδίο αφορά στο αν ζητάμε ατομική (False) ή αθροιστική (true) πιθανότητα.

β) Οι ζητούμενες αθροιστικές πιθανότητες, των ανεξάρτητων διωνυμικών ενδεχομένων, προκύπτουν εύκολα από τις ατομικές που ήδη υπολογίσαμε παραπάνω. Έτσι,
F(4) = P(X≤4) = P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4) = 0,0074 + 0,0618 + 0,2060 + 0,3433 + 0,2861 = 0,9046

P(X>2) = 1 - F(2) = 1 - {P(X=0) + P(X=1) + P(X=2)} = 1 - (0,0074 + 0,0618 + 0,2060) = 0,7248

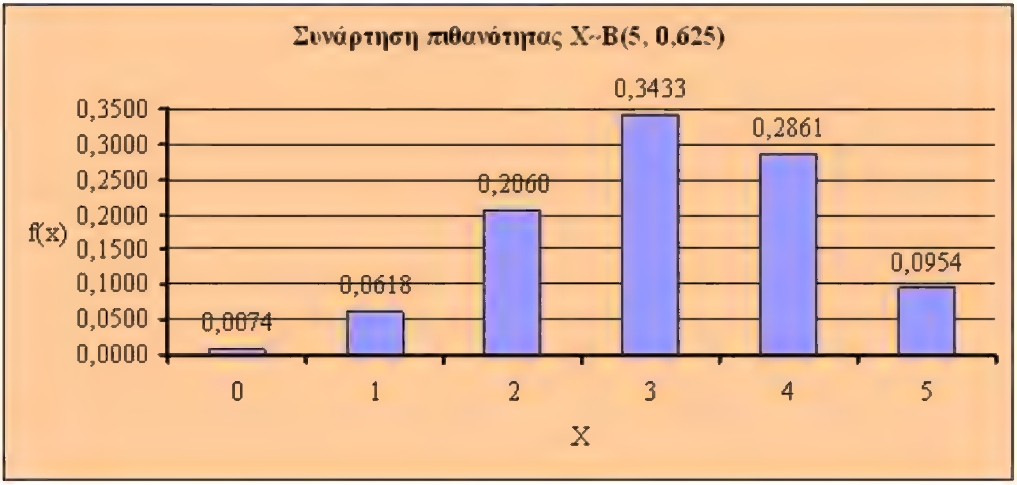
P(2≤X≤4) = F(4) – F(1) = 0,9046 - {P(X=0) + P(X=1)} = 0,9046 – 0,0692 = 0,8354

Χρησιμοποιώντας την παραπάνω συνάρτηση του MS-Excel για το «cumulative» όταν θέσουμε, αφενός «false», παίρνουμε για όλες τις τιμές της διωνυμικής μεταβλητής τη συνάρτηση πιθανότητας, και αφετέρου «true», παίρνουμε την αθροιστική ή συνάρτηση κατανομής της. Αυτές δίνονται παρακάτω τόσο σε μορφή πίνακα όσο και διαγραμματικά.

Πίνακας 6.1 Κατανομές πιθανότητας και αθροιστική ή κατανομή της X~B(5, 0,625)

X	f(x)	F(x)
0	0,0074	0,0074
1	0,0618	0,0692
2	0,2060	0,2752
3	0,3433	0,6185
4	0,2861	0,9046
5	0,0954	1,0000

Διάγραμμα 6.1 Συνάρτηση πιθανότητας



Διάγραμμα 6.2 Συνάρτηση Κατανομής

6.2.3 Παράμετροι και μορφή της διωνυμικής κατανομής

Για τον υπολογισμό των βασικών παραμέτρων, που περιγράφουν πλήρως τη διωνυμική, δηλ. μέσος μ, διακύμανση σ^2 , τυπική απόκλιση σ_x , συντελεστές ασυμμετρίας β_1 και κύρτωσης β_2 , χρησιμοποιούνται οι παρακάτω σχέσεις:

$\mu_X = E(X) = n \cdot p$	**(6.3)**
$\sigma_x^2 = Var(X) = n \cdot p \cdot q$	**(6.4)**
$\sigma_x = \sqrt{n \cdot p \cdot q}$	**(6.5)**
$\beta_1 = \dfrac{(q-p)^2}{n(pq)}$	**(6.6)**
$\beta_2 = 3 + \dfrac{1-n(pq)}{n(pq)}$	**(6.7)**

Από τις σχέσεις αυτές φαίνεται ότι, τόσο ο μέσος, όσο και η διακύμανση της διωνυμικής αυξάνουν ανάλογα με το πλήθος επαναλήψεων (n) του τυχαίου πειράματος.

Αντίθετα, όσο αυξάνει το n τόσο μειώνεται η ασυμμετρία των διωνυμικών κατανομών, ενώ εάν p=q=1/2 τότε $\beta_1=0$, δηλ. πλήρως συμμετρική με άξονα συμμετρίας την αναμενόμενη τιμή της μ.

Επίσης, η διωνυμική γίνεται μεσόκυρτη ($\beta_2=3$) όπως είναι η τυπική κανονική που θα εξετάσουμε στο επόμενο τμήμα, όσο n -> ∞ .

Παράδειγμα 6.2

Για τα δεδομένα του προηγούμενου παραδείγματος, υπολογίστε όλες τις τιμές των βασικών παραμέτρων για να περιγράψετε τη διωνυμική Χ που παριστά την επιλογή κόκκινου σφαιριδίου από κάλπη με 5 κόκκινα και 3 κίτρινα από επανάληψη της λήψης 5 φορές.

Απάντηση:

Εφαρμόζοντας τις σχέσεις (6.3-6.7) έχουμε:

- $\mu_X = E(X) = n\cdot p = 5\cdot 0,625 = 3,125$, δηλ. σε πολλές επαναλήψεις του πειράματος (ανά πεντάδα) η τάση είναι 3 στα 5 σφαιρίδια που βγάζουμε από την κάλπη (με επανάθεση) να είναι κόκκινα.

- $\sigma_X^2 = Var(X) = n\cdot p\cdot q = 5\cdot 0,625\cdot 0,375 = 1,1719$ από όπου η τυπική απόκλιση θα είναι $\sigma_X = \sqrt{1,1719} = 1,0825$ η οποία σημαίνει μέτρια (περίπου το 1/3 του μέσου) διασπορά των τιμών της υπόψη διωνυμικής γύρω από το μέσο τους μ =3,125.

- $\beta_1 = \frac{(q-p)^2}{n(pq)} = \frac{(0,375-0,625)^2}{5(0,625\cdot 0,375)} = 0,0533$ που δείχνει πολύ κοντά στη συμμετρική ($\beta_1=0$) όπως η τυπική κανονική που θα δούμε παρακάτω.

- $\beta_2 = 3 + \frac{1-n(pq)}{n(pq)} = 3 + \frac{1-5(0,625\cdot 0,375)}{5(0,625\cdot 0,375)} = 2,8533$ επίσης πολύ κοντά στο 3 που είναι η τιμή του υπόψη συντελεστή κύρτωσης (β_2) της μεσόκυρτης που είναι η τυπική κανονική.

Η μορφή ή σχήμα της υπόψη διωνυμικής, την οποία περιγράψαμε με τις τιμές των βασικών της παραμέτρων, φαίνεται παραστατικά και από το Διάγραμμα 6.1.

Από το τελευταίο παράδειγμα, αλλά κυρίως από τις σχέσεις (6.3-6.7), ο προσεκτικός αναγνώστης πρέπει να έχει αντιληφθεί ότι το σχήμα του γραφήματος της διωνυμικής

συνάρτησης πιθανότητας εξαρτάται κυρίως από το πλήθος των επαναλήψεων (n) του πειράματος και την πιθανότητα «επιτυχίας» (p). Πιο συγκεκριμένα:

α) αν p = 0,50 η διωνυμική είναι συμμετρική με μορφή κάθετης τομής αντίστροφης καμπάνας, όπως ο ακρογωνιαίος λίθος της Στατιστικής, η τυπική κανονική που θα περιγράψουμε παρακάτω.

β) αν p > 0,50 και n μικρός αριθμός, η κατανομή έχει αρνητική ασυμμετρία (μεγαλύτερη η αριστερή της δεξιάς ουράς), η οποία αυξάνεται όσο η πιθανότητα επιτυχίας τείνει στη μονάδα.

γ) αν p < 0,50 και n μικρός αριθμός, η κατανομή έχει θετική ασυμμετρία, η οποία μειώνεται όσο η πιθανότητα επιτυχίας τείνει στο μηδέν.

δ) για δεδομένη πιθανότητα επιτυχίας η διωνυμική κατανομή τείνει στη μεσόκυρτη συμμετρική όπως αυτή της τυπικής κανονικής, όσο αυξάνει το πλήθος των επαναλήψεων (n) του πειράματος.

Παράδειγμα 6.3

Τα αποτελέσματα έρευνας που αφορούσε τον αριθμό των γυναικών στο ανθρώπινο δυναμικό του τμήματος προμηθειών εμπορικών επιχειρήσεων στην Ελλάδα δίνονται στον παρακάτω πίνακα. Το τυχαίο δείγμα κάλυψε 368 εμπορικές επιχειρήσεις που απασχολούσαν μέχρι και 6 γυναίκες στο υπόψη τμήμα. Μπορούμε να ισχυριστούμε ότι η πιθανοθεωρητική συμπεριφορά των επιχειρήσεων του υπόψη κλάδου ακολουθεί το διωνυμικό νόμο πιθανότητας;
Κατανομή Συχνοτήτων

xi	Fi
0	6
1	30
2	82
3	120
4	97
5	25
6	8
Σύνολο	368

Απάντηση:

Αυτό που πραγματικά ζητείται είναι να ελέγξουμε, με περιγραφικά στατιστικά κριτήρια, εάν στη δεδομένη εμπειρική κατανομή συχνότητας προσαρμόζεται ικανοποιητικά η διωνυμική κατανομή πιθανότητας.

Για το σκοπό αυτό αρκεί να υπολογίσουμε τις θεωρητικές συχνότητες που θα έχει η υπόψη κατανομή εάν πράγματι ακολούθησε τη διωνυμική. Οι θεωρητικές συχνότητες μπορούν να υπολογιστούν εύκολα χρησιμοποιώντας τον ορισμό της εμπειρικής πιθανότητας. Έτσι

σε πρώτο στάδιο, πρέπει να υπολογίσουμε τις ατομικές πιθανότητες για κάθε μία τιμή της τυχαίας μεταβλητής X, κάτω από την υπόθεση ότι αυτή ακολουθεί το γνωστό θεωρητικό νόμο της διωνυμικής.

Από τα δεδομένα έχουμε ότι το μέγεθος του τυχαίου δείγματος είναι N=368, n=6, ενώ ορίζουμε ως επιτυχία «γυναίκα εργαζόμενη στο τμήμα προμηθειών».

Υπολογίζουμε τη σταθερή πιθανότητα p επιτυχίας χρησιμοποιώντας τον τύπο (6.3) της αναμενόμενης τιμής της διωνυμικής. Επομένως, καταρχήν υπολογίζουμε τον εμπειρικό μέσο αριθμητικό:

$$\bar{X} = \frac{\sum_i f_i x_i}{N} = \frac{1.115}{368} = 3,03$$

Αφού έχουμε υποθέσει ότι η υπόψη εμπειρική ακολουθεί τη διωνυμική μπορούμε να γράψουμε:

$$\mu_X = E(X) = n \cdot p \Rightarrow 3,03 = 6 \cdot p \Rightarrow p = 0,5050$$

Στη συνέχεια υπολογίζουμε τις διωνυμικές πιθανότητες κατά τα γνωστά, π.χ. με τον αναδρομικό τύπο.

$$p_0 = q^n = 0,4950^6 = 0,0147$$

$$p_{x+1} = \frac{n-x}{x+1} \cdot \frac{p}{q} \cdot p_x \Leftrightarrow p_{0+1} = \frac{6-0}{0+1} \cdot 1,0201 \cdot 0,0147 \Rightarrow p_1 = 0,0900$$

κ.λπ. Οι υπολογισμοί δίνονται στον παρακάτω Πίνακα.

Πίνακας 6.2 Υπολογισμοί για την προσαρμογή διωνυμικής σε δοσμένη εμπειρική

Εμπειρική Κατανομή Συχνοτήτων			Διωνυμικές πιθανότητες	Θεωρητικές συχνότητες
x_i	f_i	$f_i x_i$	$f(x_i)$	$\Theta_i = N \cdot f(x_i)$
0	6	0	0,0147	5
1	30	30	0,0900	33
2	82	164	0,2297	85
3	120	360	0,3124	115
4	97	388	0,2390	88
5	25	125	0,0975	36
6	8	48	0,0166	6
Σύνολο	N =368	1.115	1,0000	368

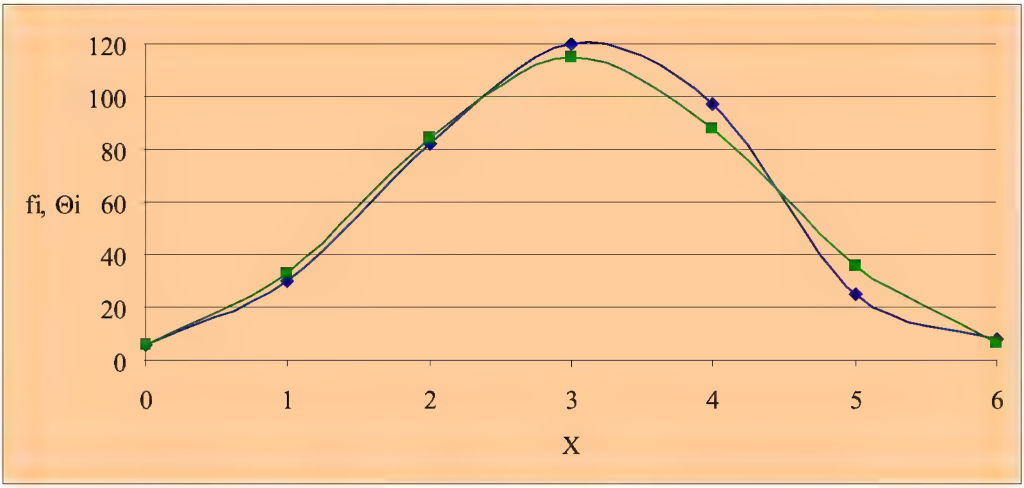
Οι διωνυμικές πιθανότητες μπορούν να υπολογιστούν από οποιοδήποτε λογισμικό όπως πχ. με το MS-Excel που δείξαμε προηγουμένως.

Όπως φαίνεται από τη σύγκριση εμπειρικών (f_i) και θεωρητικών, εδώ διωνυμικών, (Θ_i) συχνοτήτων δεν υπάρχουν «σημαντικές» διαφορές. Αυτό φαίνεται και από το παρακάτω διάγραμμα.

Εντούτοις, οι υπόψη διαφορές (f_i-Θ_i) μπορεί να ελεγχθούν κατά πόσο είναι «στατιστικά σημαντικές», δηλ. δεν οφείλονται στις κυμάνσεις της τυχαίας δειγματοληψίας και επομένως το συμπέρασμα μας σε όρους πιθανότητας να μπορεί να γενικευθεί με δεδομένη ακρίβεια. Ο έλεγχος αυτός είναι ένας μη-παραμετρικός χ^2 που δεν εμπίπτει στην ύλη του παρόντος κεφαλαίου.

Κατά συνέπεια τα όποια συμπεράσματα προκύπτουν από την παρούσα ανλυση, αν και λαμβάνουν υπόψη τους τη διωνυμική κατανομή πιθανότητας για τον υπολογισμό των θεωρητικών συχνοτήτων, δεν συνιστούν στατιστική συμπερασματολογία με την έννοια της επαγωγής στον πληθυσμό, αλλά αντίστοιχα μόνο περιγραφική.

<u>Διάγραμμα 6.3</u> Προσαρμογή διωνυμικής σε εμπειρική κατανομή συχνότητας



6.3 Κανονική κατανομή

6.3.1 Η έννοια της κανονικής κατανομής

Ακρογωνιαίος λίθος της Στατιστικής συμπερασματολογίας είναι η κατανομή των Gauss-Laplace ή κανονική κατανομή μιας συνεχούς τυχαίας μεταβλητής X με παραμέτρους μ και σ^2 με συνάρτηση πυκνότητας πιθανότητας η οποία δίνεται από την παρακάτω σχέση:

$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < +\infty$	**(6.8)**

Όπου π=3,14159... και e=2,71828...
Όπως πάντα, για να είναι η σχέση αυτή συνάρτηση πυκνότητας πιθανότητας πρέπει να ικανοποιούνται οι γνωστές από το προηγούμενο κεφάλαιο προϋποθέσεις, των θετικών οριακών πιθανοτήτων και του βέβαιου ενδεχόμενου του δειγματικού χώρου,

$$f(x)\geq 0 \quad ,-\infty < x < +\infty \quad \kappa\alpha\iota \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

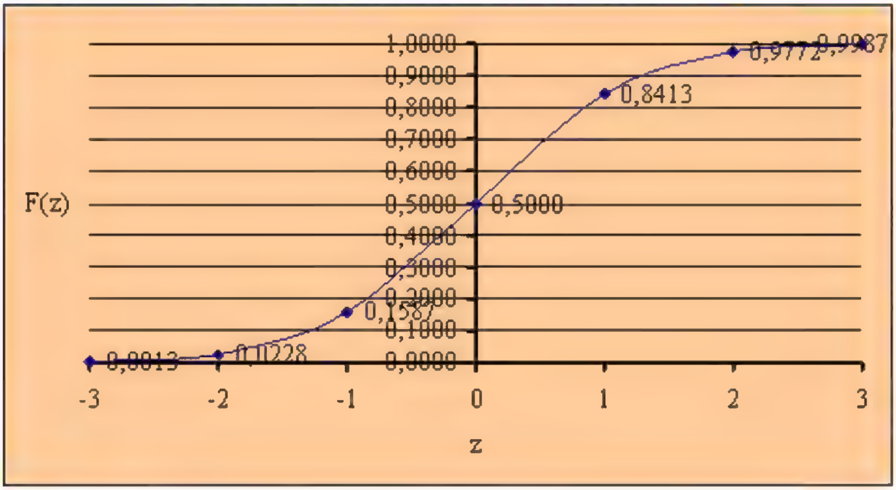
6.3.2 Ιδιότητες της κανονικής κατανομής

Όπως φαίνεται από τη συνάρτηση πυκνότητας η κανονική ορίζεται πλήρως από το μέσο και τη διακύμανσή της, δηλ. από το ζεύγος (μ, σ^2). Έχει μορφή αντεστραμμένου κώδωνα και είναι συμμετρική με άξονα συμμετρίας την τιμή x=μ. Το σημείο αυτό είναι και το μέγιστο, δηλ. $f(x)=\frac{1}{\sigma\sqrt{2\pi}}$. Επομένως, ο μέσος αφενός, συμπίπτει με τη διάμεσο και τον τύπο, και αφετέρου, χωρίζει το εμβαδόν κάτω από την κανονική καμπύλη σε δύο ίσα μέρη (βλ. παρακάτω διάγραμμα συνάρτησης πυκνότητας).

Οι πίνακες της συνεχούς κανονικής συνάρτησης πιθανότητας είναι αθροιστικοί και δίνουν πιθανότητες της μορφής P(Z<z), μόνο για μία κανονική τυχαία μεταβλητή την τυπική (standard normal distribution). Η τελευταία είναι η κανονική με μέσο μηδέν και διακύμανση τη μονάδα, δηλ. Z~N(μ=0, σ^2=1).

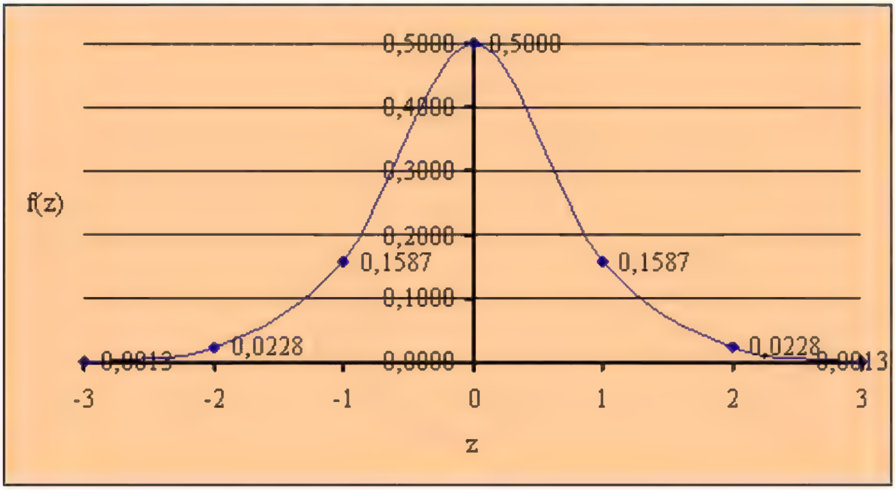
Το γράφημα της συνάρτησης κατανομής της τυπικής κανονικής φαίνεται στο παρακάτω διάγραμμα.

Διάγραμμα 6.4 **Συνάρτηση κατανομής τυπικής κανονικής**



Εξάλλου, το γράφημα της συνάρτησης πυκνότητας της τυπικής, που δίνει το εμβαδόν κάτω από την καμπύλη φαίνεται στο παρακάτω διάγραμμα.

Διάγραμμα 6.5 **Συνάρτηση πυκνότητας πιθανότητας τυπικής κανονικής**



Εντούτοις, το ενδιαφέρον μας δεν εστιάζεται στην ειδική περίπτωση της Z τυπικής κανονικής κατανομής, αλλά στην απειρία των κανονικών X τυχαίων μεταβλητών με παραμέτρους (μ, σ^2). Έτσι για να μεταφέρουμε τις γνωστές πιθανότητες των Gauss-Laplace f(z) στις άγνωστες αλλά ζητούμενες κανονικές f(x) χρησιμοποιούμε τον λεγόμενο τυπικό μετασχηματισμό που δίνει τη σχέση μεταξύ X και Z, δηλ.:

$$Z = \frac{X - \mu}{\sigma} \qquad \textbf{(6.9)}$$

Κατά συνέπεια, η συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής δίνεται από τη σχέση:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}, \quad -\infty < z < +\infty \qquad \textbf{(6.10)}$$

με τις γνωστές προϋποθέσεις των συναρτήσεων πιθανότητας:

$$f(z) \geq 0, \ -\infty < z < +\infty \ \text{και} \ \int_{-\infty}^{+\infty} f(z)dz = 1$$

6.3.3 Πιθανότητες και κρίσιμες τιμές κανονικών τυχαίων μεταβλητών

Μία από τις σημαντικότερες συνέπειες της συμμετρικότητας της κανονικής κατανομής είναι ότι το εμβαδόν κάτω από την καμπύλη σε 3 ή 2 ή 1 ή k τυπικές αποκλίσεις από το μέσο είναι το ίδιο για οποιαδήποτε κανονική X, δηλ. έχουν όλες την ίδια πιθανότητα P(|X-μ|<kσ). Με άλλα λόγια, η απάντηση στο ερώτημα «ποια είναι η πιθανότητα η κανονική τυχαία X να παίρνει τιμές στα διαστήματα μ±3σ , μ±2σ , μ±σ κ.λπ.» ή εναλλακτικά «ποιο ποσοστό των τιμών της X βρίσκεται στα διαστήματα μ±3σ , μ±2σ , μ±σ κ.λπ.» δίνεται παρακάτω (βλ. διάγραμμα 6.5) από τους πίνακες της τυπικής κανονικής.

Εντούτοις, είναι προφανές ότι αυτές ισχύουν για οποιαδήποτε X κανονική {P(|X-μ|<kσ)= P(|Z|<k)=P(-k<Z<k)=F(k)-F(-k), k=1,2,...}

- P(-3<Z<3)=F(3)-F(-3)=0,9987-0,0013=0,9973

ή υπάρχει πιθανότητα 99,73% τυχαία επιλεγόμενη τιμή z_i να προέρχεται από το διάστημα μ±3σ (μ=0 και σ=1, εδώ) ή ακόμα ότι το 99,73% των τιμών της X βρίσκονται στο διάστημα μ±3σ.

- P(-2<Z<2)=F(2)-F(-2)=0,9772-0,0228=0,9545

ή υπάρχει πιθανότητα 95,45% τυχαία επιλεγόμενη τιμή z_i να προέρχεται από το διάστημα μ±2σ (μ=0 και σ=1, εδώ) ή ακόμα ότι το 95,45% των τιμών της X βρίσκονται στο διάστημα μ±2σ.

- P(-1<Z<1)=F(1)-F(-1)=0,8434-0,15866=0,6827

ή υπάρχει πιθανότητα 68,27% τυχαία επιλεγόμενη τιμή z_i να προέρχεται από το διάστημα μ±σ (μ=0 και σ=1, εδώ) ή ακόμα ότι το 68,27% των τιμών της X βρίσκονται στο διάστημα μ±σ.

Τις πιθανότητες αυτές μπορεί εύκολα να βρει ο αναγνώστης από τους Πίνακες της τυπικής κανονικής (Z) στα εγχειρίδια στατιστικής. Εκεί το κύριο σώμα του Πίνακα δίνει αθροιστικές πιθανότητες ενώ η πρώτη γραμμή και πρώτη στήλη συνθέτουν από κοινού τις τιμές (z_i) μέχρι τις οποίες δίνονται αυτές. Έτσι για παράδειγμα αν ενδιαφερόμαστε για την P(Z<1,48), τότε ανατρέχοντας στον υπόψη στατιστικό Πίνακα των Gauss-Laplace κοιτάμε στη γραμμή που έχει τίτλο «1,4» δηλ. στη γραμμή με το πρώτο δεκαδικό ψηφίο της z_i που ζητείται, και στη στήλη «0,08» δηλ. στη στήλη με το δεύτερο δεκαδικό ψηφίο της z_i που ζητείται. Στη συντεταγμένη αυτή βρίσκουμε τη ζητούμενη πιθανότητα P(Z<1,48)=0,4306 (βλ. παρακάτω Πίνακα 6.3).

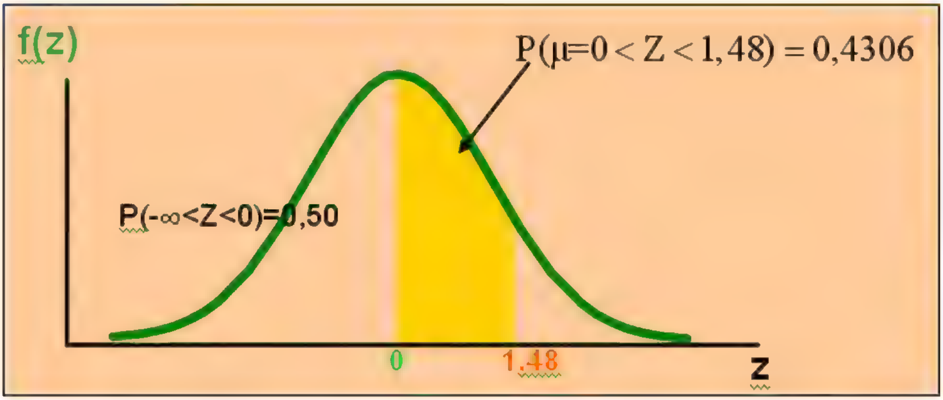
Προσοχή πρέπει να δοθεί στο εάν ο Πίνακας περιέχει τις P(-∞<Z<+∞) ή μόνο εκείνες P(0<Z<+∞). Στην τελευταία περίπτωση οι υπόλοιπες προκύπτουν αβίαστα λαμβάνοντας υπόψη τη συμμετρικότητα της κατανομής. Έτσι, στο υπόψη παράδειγμα θα έχουμε (βλ. Πίνακα 6.3):

- P(Z<1,48)=P(Z<0)+P(0<Z<1,48)=0,5000+0,4306=0,9306

Πίνακας 6.3 Τμήμα του Πίνακα αθροιστικών πιθανοτήτων της τυπικής κανονικής Z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Διάγραμμα 6.6



Οι Πίνακες που περιέχουν τις υπόψη πιθανότητες της τυπικής κανονικής, μέσω της οποίας με τον τυπικό μετασχηματισμό βρίσκουμε τις αντίστοιχες για όλες τις X~N(μ, σ^2) περιέχονται σε πολλά σχετικά λογισμικά όπως π.χ. στο MS-Excel. Στο τελευταίο η συνάρτηση κατανομής της τυπικής δίνεται από την «NORMSDIST(z_i)». Έτσι αν ψάχνουμε την P(Z<1) εισάγουμε στο όρισμα της συνάρτησης 1, δηλ. «NORMSDIST(1)» που αποδίδει «=0,8413», δηλ. το ζητούμενο P(Z<1)=0,8413.

Για την περίπτωση που χρειαζόμαστε πιθανότητες που δεν έχει ο Πίνακας αφού δίνει μόνο P(0<Z<+∞), αβίαστα αυτές προκύπτουν από τα παραπάνω και τη συμμετρικότητα της κανονικής. Για παράδειγμα:

- P(Z<-3)=P(Z>3)=F(-3)=1-F(3)=0,0013

ή υπάρχει πιθανότητα 0,013% τυχαία επιλεγόμενη τιμή z_i να προέρχεται από τα διαστήματα (-∞, -3) ή (3, +∞) ή ακόμα ότι το 0,013% των τιμών της Z βρίσκονται στα διαστήματα (-∞, -3) ή (3, +∞).

- P(Z<-2)=P(Z>2)=F(-2)=1-F(2)=0,0228

ή υπάρχει πιθανότητα 2,28% τυχαία επιλεγόμενη τιμή z_i να προέρχεται από τα διαστήματα (-∞, -3) ή (3, +∞) ή ακόμα ότι το 2,28% των τιμών της Z βρίσκονται στα διαστήματα (-∞, -3) ή (3, +∞).

- P(Z<-1)=P(Z>1)=F(-1)=1-F(1)=0,1587

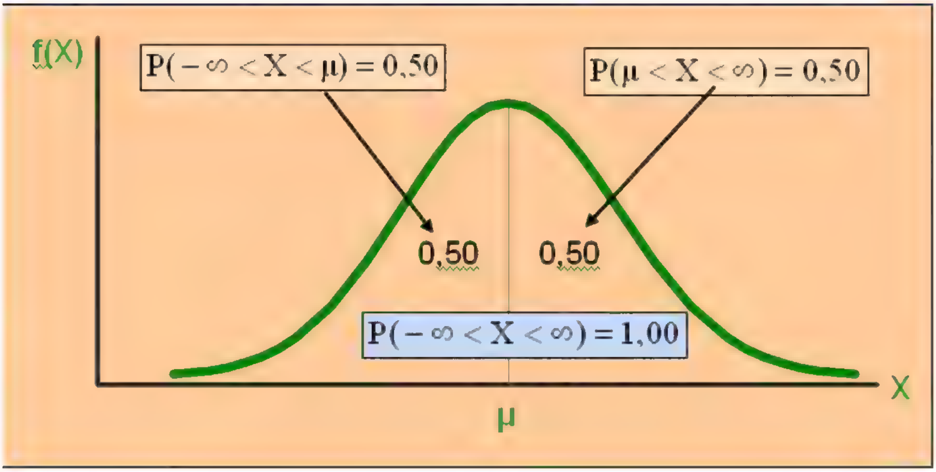
ή υπάρχει πιθανότητα 15,87% τυχαία επιλεγόμενη τιμή z_i να προέρχεται από τα διαστήματα (-∞, -3) ή (3, +∞) ή ακόμα ότι το 15,87% των τιμών της Z βρίσκονται στα διαστήματα (-∞, -3) ή (3, +∞).

Επίσης, εύκολα μπορεί να διαπιστώσει ο αναγνώστης ασκούμενος στον υπολογισμό πιθανοτήτων κάτω από την τυπική κανονική, την ισχύ των παρακάτω:

- P(-3<Z<-2)=F(-2)-F(-3)=0,0228-0,0013=0,0214 και
  P(-3<Z<-2)=F(2<Z<3)=F(3)-F(2)=0,9987-0,9772=0,0214
- P(-2<Z<-1)=F(-1)-F(-1)=0,1587-0,2228=0,1359 και
  P(-2<Z<-1)=F(1<Z<2)=F(2)-F(1)=0,9772-0,8413=0,1359

- P(-1<Z<0)=F(0)-F(-1)=0,50-0,1587=0,3413 και
  P(-1<Z<0)=F(0<Z<1)=F(1)-F(0)=0,8413-0,50=0,3413

Διάγραμμα 6.7 Συμμετρικότητα των κανονικών Χ περί το μέσο τους



Παράδειγμα 6.4

Η ηλικία εργαζομένων μεγάλης επιχείρησης ακολουθεί την κανονική κατανομή πιθανότητας με μέσο 55 και τυπική απόκλιση 9 έτη, δηλ. X~N(55, 81).

Ζητείται **α)** τι πιθανότητα υπάρχει τυχαία επιλεγόμενος εργαζόμενος να έχει ηλικία μεταξύ 46 και 60 ετών; **β)** τι πιθανότητα υπάρχει τυχαία επιλεγόμενος εργαζόμενος να έχει ηλικία μέχρι 58 έτη; **γ)** τι πιθανότητα υπάρχει τυχαία επιλεγόμενος εργαζόμενος να είναι πάνω από 62 ετών; **δ)** πόσοι έχουν ηλικία κάτω από 50 ετών, αν γνωρίζετε ότι η επιχείρηση απασχολεί 300 εργαζομένους;

Απάντηση:

Χρησιμοποιούμε τον τυπικό μετασχηματισμό για να μετατρέψουμε τις x_i σε z_i τις αθροιστικές πιθανότητες των οποίων έχουμε σε πίνακα. Έτσι, θα έχουμε:

α)

$$P(46 < X < 60) = P(\frac{46-55}{9} < \frac{X-\mu}{\sigma} < \frac{60-55}{9}) = P(-1 < Z < 0,5556)$$
$$= F_Z(0,5556) - F_Z(-1) = 0,7108 - 0,1587 = 0,5521$$

Επομένως, υπάρχει πιθανότητα 55,21% τυχαία επιλεγόμενος εργαζόμενος να έχει ηλικία μεταξύ 46 και 60 ετών.

β)

$$P(X<58)=P(\frac{X-\mu}{\sigma}<\frac{58-55}{9})=P(Z<0,3333)=F_Z(0,3333)=0,6305$$

Επομένως, υπάρχει πιθανότητα 63,05% τυχαία επιλεγόμενος εργαζόμενος να έχει ηλικία μέχρι 48 ετών.

γ)

$$P(X>62)=1-P(X<62)=1-P(\frac{X-\mu}{\sigma}<\frac{62-55}{9})=1-P(Z<0,7778)$$

$$=1-F_Z(0,7778)=1-0,7817=0,2183$$

Επομένως, υπάρχει πιθανότητα 21,83% τυχαία επιλεγόμενος εργαζόμενος να έχει ηλικία μεγαλύτερη των 62 ετών.

δ) Υπολογίζουμε την πιθανότητα P(X<50), την οποία στη συνέχεια πολλαπλασιάζουμε με το μέγεθος του δείγματος N=300 για να βρούμε τη θεωρητική απόλυτη συχνότητα που ζητείται. Εφαρμόζουμε δηλ. την έννοια της εμπειρικής πιθανότητας, που ως γνωστόν είναι το όριο της σχετικής συχνότητας.

$$P(X<50)=P(\frac{X-\mu}{\sigma}<\frac{50-55}{9})=P(Z<-0,5556)=F_Z(-0,5556)=0,2892$$

Επομένως, περίπου 87 εργαζόμενοι (=300·0,2892) στους 300 της επιχείρησης έχουν ηλικία μέχρι 50 ετών.

7. Εκπαιδευτική Ενότητα

- **Δειγματοληψία και Κατανομές Δειγματοληψίας**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να αναγνωρίζει και να ορίζει το τυχαίο δείγμα, τα στατιστικά σφάλματα, τα είδη δειγμάτων, τα δειγματοληπτικά σχέδια.
- Να γνωρίζει και να εφαρμόζει τις κατανομές δειγματοληψίας καθώς και την έννοια και χρησιμότητα κατανομών δειγματοληψίας.
- Να εφαρμόζει και να υλοποιεί τις κατανομές βασικών στατιστικών: μέσου, τυπικής απόκλισης και ποσοστού ή αναλογίας.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Τυχαίο δείγμα
- Στατιστικά σφάλματα
- Είδη δειγμάτων
- Δειγματοληπτικά σχέδια
- Κατανομές δειγματοληψίας

7.1 Εισαγωγή

Η μελέτη των στοιχείων του στατιστικού πληθυσμού όχι μόνο είναι χρονοβόρα, και απαγορευτική από άποψη κόστους, τις περισσότερες φορές, αλλά υπάρχουν περιπτώσεις στις οποίες δεν είναι δυνατόν να είναι διαθέσιμα τα στατιστικά στοιχεία, όπως για παράδειγμα όταν πρόκειται για άπειρους ή άγνωστους πληθυσμούς.

Αναγκαστικά λοιπόν συνήθως γεννιέται η ανάγκη μελέτης δειγμάτων, τα οποία χρησιμοποιούμε για να εκτιμήσουμε τα χαρακτηριστικά του γεννήτορα αλλά αγνώστου πληθυσμού που μας ενδιαφέρει. Δείγμα, ορίζουμε ένα υποσύνολο του συνόλου των παρατηρήσεων του πληθυσμού. Το υποσύνολο όμως αυτό επιλέγεται με ορισμένους κανόνες που έχει θεμελιώσει η στατιστική επιστήμη στον ιδιαίτερο κλάδο της, τη δειγματοληψία. Οι κανόνες αυτοί ορίζουν τις τεχνικές δειγματοληψίας (sampling techiques) ή σχέδια δειγματοληψίας (sampling designs) με βάση τα οποία, επιλέγονται οι δειγματοληπτικές μονάδες που έχουν ιδιαίτερα χαρακτηριστικά, ταυτοποιώντας το δείγμα που τελικά θα μελετήσουμε, σε δύο κύριες καταρχήν κατηγορίες των τυχαίων και μητυχαίων δειγμάτων. Βασικά σχέδια δειγματοληψίας, ειδικά για τη στατιστική επαγωγή που μελετάμε στα επόμενα κεφάλαια, παρουσιάζουμε στο δεύτερο τμήμα αυτού του κεφαλαίου.

Από θεωρητικής άποψης οι στατιστικολόγοι στη δειγματοληψία έχουν θεμελιώσει τις ιδιότητες των χαρακτηριστικών του δείγματος με δεδομένες (υποθετικά γνωστές) τις παραμέτρους του πληθυσμού. Η μεθοδολογία αυτή συνιστά τη στατιστική επαγωγή (staitical deduction), η οποία είναι η βάση και για την (αντίστροφη προσέγγιση) εξαγωγή συμπερασμάτων για τον πληθυσμό, από τα δεδομένα του δείγματος που διερευνούμε. Η τελευταία προσέγγιση, δηλαδή από το δείγμα στον πληθυσμό, είναι η στατιστική επαγωγή (statistical induction) ή στατιστική συμπερασματολογία (statistical inference) που είναι το αντικείμενο των επόμενων κεφαλαίων.

Οι υπόψη ιδιότητες των χαρακτηριστικών του δείγματος, για να είναι αυτό αντιπροσωπευτικό της δομής του πληθυσμού, έτσι ώστε να μπορούμε να γενικεύσουμε τα συμπεράσματά του στο γεννήτορα πληθυσμό, περιγράφονται αναλυτικά από τις λεγόμενες κατανομές δειγματοληψίας, δηλαδή από τις κατανομές πιθανότητας των στατιστικών (συναρτήσεων ή απλά στατιστικών) του δείγματος, π.χ. του μέσου, της διακύμανσης, της τυπικής απόκλισης, της αναλογίας κ.λ.π. Ορισμένες ευρέως χρησιμοποιούμενες από αυτές τις κατανομές δειγματοληψίας παρουσιάζουμε στο τρίτο τμήμα αυτού του κεφαλαίου.

Επισημαίνεται ότι, όπως στα προηγούμενα αλλά και στα επόμενα κεφάλαια, η ανάλυσή μας γίνεται σε επίπεδο εφαρμοσμένης επαγωγικής στατιστικής, με την έννοια ότι τα στοιχεία θεωρίας που παρουσιάζουμε αποτελούν απλώς «αναφορά», αν και συστηματική, εντούτοις όμως λιτή, που σημαίνει ότι ο ενδιαφερόμενος αναγνώστης πρέπει να ανατρέξει στα σχετικά εγχειρίδια στατιστικής συμπερασματολογίας που ενδεικτικά παρουσιάζονται στη βιβλιογραφία. Η έμφαση εδώ δίνεται στις εφαρμογές.

7.2 Περί δειγματοληψίας

7.2.1 Η έννοια του τυχαίου δείγματος

Αποδεικνύεται ότι αν το δείγμα μεγέθους (n) που μελετάμε δεν είναι τυχαίο, με τη στατιστική έννοια του όρου, τότε δεν μπορούμε να χρησιμοποιήσουμε την πιθανοθεωρία για να επαγάγουμε (γενικεύσουμε) τα συμπεράσματα για τα χαρακτηριστικά του, στο γεννήτορα άγνωστο πληθυσμό, ο οποίος όμως είναι ο σκοπός μας. Έτσι, είναι προφανής η αξία της επιλογής τυχαίων δειγμάτων, για την αξιοπιστία της στατιστικής συμπερασματολογίας.

Απλό τυχαίο ή τυχαίο δείγμα μεγέθους (n) παρατηρήσεων από πληθυσμό μεγέθους N, για την τυχαία μεταβλητή X, ονομάζεται αυτό που έχει ως ιδιότητες, αφενός, ότι κάθε μία από τις n παρατηρήσεις του έχει την ίδια και γνωστή εκ των προτέρων πιθανότητα επιλογής σε αυτό, και αφετέρου, ότι όλα τα δυνατά ισομεγέθη δείγματα n από τον πληθυσμό N έχουν και αυτά την ίδια και γνωστή εκ των προτέρων πιθανότητα επιλογής.

Επομένως, οι λήψεις των δειγματοληπτικών μονάδων πρέπει να είναι ανεξάρτητες, κάτι που ορίζει και τις τιμές x_i τυχαίες και ανεξάρτητες παρατηρήσεις, ενώ και η κατανομή πιθανότητας του πληθυσμού παραμένει σταθερή από λήψη σε λήψη. Προϋπόθεση δηλ. του τυχαίου δείγματος είναι η δειγματοληψία να γίνεται, είτε, με επανατοποθέτηση κάθε επιλεγόμενης παρατήρησης (δειγματοληπτική μονάδα) στη μήτρα του πληθυσμού, είτε, από άπειρο πληθυσμό.

Με τα τυχαία δειγματοληπτικά σχέδια, επιτυγχάνεται ο κύριος σκοπός της δειγματοληψίας που είναι η επιλογή αντιπροσωπευτικών δειγμάτων της δομής του πληθυσμού από τον οποίο προέρχονται. Μόνο από τυχαία δείγματα μπορούμε να βγάλουμε αξιόπιστα, και μετρήσιμα σε όρους πιθανότητας, συμπεράσματα για τα άγνωστα αλλά αληθή χαρακτηριστικά του πληθυσμού, τα οποία εκτιμούμε από τα δεδομένα των τυχαίων αυτών δειγμάτων.

Με αυστηρή διατύπωση, αν από πληθυσμό με συνάρτηση πυκνότητας f(x) επιλέγεται δείγμα μεγέθους n, οι παρατηρήσεις του οποίου προέρχονται από τις μεταβλητές X_1, X_2, ..., Xn , θα λέμε ότι αυτό είναι τυχαίο αν οι X_1, X_2, ..., X_n κατανέμονται αμοιβαία ανεξάρτητα και έχουν όλες την ίδια ακριβώς κατανομή (identically distributed). Για να ικανοποιούνται οι προϋποθέσεις της ανεξαρτησίας και της κοινής κατανομής πιθανότητας των δειγματοληπτικών μονάδων, απαιτείται να δειγματοληπτούμε από άπειρο πληθυσμό, ή από πεπερασμένο με επανάθεση.

Επομένως, το δείγμα με στοιχεία τις τυχαίες μεταβλητές X_1, X_2, ..., X_n θα έχει από κοινού συνάρτηση πιθανότητας (join probability function) την, έστω,$\gamma(X_1, X_2, ..., X_n)$ η οποία ονομάζεται κατανομή του δείγματος (sample's join probability distribution). Αφού οι παρατηρήσεις είναι ανεξάρτητες, η κατανομή του δείγματος θα ισούται με το γινόμενο των οριακών (marginal) συναρτήσεων πυκνότητας πιθανότητάς τους. Με άλλα λόγια, για την κατανομή του δείγματος ισχύει η σχέση:

$$\gamma(x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n) = \prod_i f(x_i) \qquad (7.1)$$

Τονίζεται ιδιαίτερα ότι, στη συνάρτηση πιθανότητας του δείγματος (7.1) οι τιμές x_i, (i=1,2,...,n) του δείγματος θεωρούνται ως μεταβλητές, αντίθετα με τις πληθυ-σμιακές παραμέτρους π.χ. μέσος, διακύμανση κ.λπ. ή γενικά θ_j, (j=1,2,...,λ) οι οποίες εδώ θεωρούνται δεδομένες και σταθερές.

Αξιοσημείωτο είναι ότι στο τυχαίο δείγμα ο όρος τυχαίο αναφέρεται στον τρόπο συλλογής των στοιχείων (x_i), δηλ. δεν είναι ιδιότητα των παρατηρήσεων, αλλά της δειγματοληπτικής διαδικασίας (τυχαία δειγματοληπτικά σχέδια), η οποία εξασφαλίζει την τυχαιότητα επιλογής των ανεξάρτητων, και με κοινή συνάρτηση πιθανότητας, δειγματοληπτικών μονάδων.

Παράδειγμα 7.1

Έστω ότι Χ τυχαία μεταβλητή με κατανομή πιθανότητας:

$$f(x)=\begin{cases} 1/4 & \text{για} \quad x=0 \\ 2/4 & \text{για} \quad x=1 \\ 1/4 & \text{για} \quad x=2 \\ 0 & \text{για} \quad x\neq 0,1,2 \end{cases}$$

Ποια είναι η πιθανότητα να επιλέξουμε το δείγμα (1,1,2) από τον πληθυσμό αυτό (Χ);

Απάντηση:

Αφού κάθε παρατήρηση του πληθυσμού είναι τυχαία μεταβλητή, εφαρμόζοντας την από κοινού συνάρτηση πυκνότητας πιθανότητας (7.1) για το ζητούμενο δείγμα (1,1,2) βρίσκουμε ότι αυτό έχει πιθανότητα επιλογής 6,25%, όπως φαίνεται παρακάτω

$$\gamma(1,1,2)=f(1)\cdot f(1)\cdot f(2)=(2/4)(2/4)(1/4)=4/64$$

Με άλλα λόγια η πιθανότητα επιλογής του δείγματος (1,1,2), από τον πληθυσμό που παριστά η Χ με κατανομή πιθανότητας που δίνεται στην εκφώνηση της άσκησης, είναι 4/64=0,0625.

Αυτό όμως που εφαρμόζουμε συνήθως στην κλασική επαγωγική στατιστική συμπερασματολογία είναι να πιθανολογούμε για το γεννήτορα πληθυσμό με βάση τις δεδομένες τιμές του δείγματος. Εφαρμόζουμε, δηλ. αντίστροφη μεθοδολογία από αυτήν της συνάρτησης του δείγματος (7.1). Η προσπάθεια αυτή διευκολύνεται ιδιαίτερα από τη λεγόμενη συνάρτηση πιθανοφάνειας του δείγματος (likelihood function), σύμφωνα με την οποία οι τιμές X_i, (i=1,2,...,n) του τυχαίου μας δείγματος είναι γνωστές (σταθερές), ενώ είναι τώρα οι πληθυσμιακοί παράμετροι, έστω, θ_j, (j=1,2,...,λ) άγνωστες.

Με άλλα λόγια, η συνάρτηση πιθανοφάνειας μας δίνει πιθανότητες να πάρουμε τις παρατηρήσεις του δείγματος μας, εφόσον οι άγνωστες παράμετροι έχουν ορισμένες τιμές, στο εύρος που κυμαίνονται αφού ως άγνωστες θεωρούνται μεταβλητές. Με βάση τα σύμβολα που χρησιμοποιούμε, η συνάρτηση πιθανοφάνειας του δείγματος δίνεται από τη σχέση:

$L(x_1, x_2, \ldots, x_n) = \prod_{i,j} f(x_i \mid \theta_j),$ $i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, \lambda$	**(7.2)**

Παράδειγμα 7.2

Έστω ότι Χ διωνυμική μεταβλητή με κατανομή πιθανότητας:

$$f(x \mid p, n=5) = \begin{cases} p & \text{για} \quad x = 1 \\ 1-p & \text{για} \quad x = 0 \end{cases}, \; 0 \le p \le 1 \text{ και } f(x \mid p, n=5) = 0 \text{ για } x \neq 0,1$$

Ποια είναι η πιθανότητα να έχουμε το δείγμα (1,1,0) από τον πληθυσμό αυτό (Χ);

Απάντηση:

Εφαρμόζοντας τη συνάρτηση πιθανοφάνειας (7.2) θα έχουμε:

$$L(x_1, x_2, \ldots, x_n) = \prod_{i,j} f(x_i \mid \theta_j)$$

$$= L(1,1,0) = f(1 \mid p) \cdot f(1 \mid p) \cdot f(0 \mid 1-p) = p \cdot p \cdot (1-p) = p^2 \cdot (1-p)$$

Βλέπουμε δηλ. ότι η πιθανοφάνεια του δείγματος (1,1,0) είναι συνάρτηση της παραμέτρου p. Άρα, για ορισμένες ενδεικτικές τιμές της διωνυμικής πιθανότητας επιτυχίας p θα έχουμε αντίστοιχες πιθανότητες για τις σταθερές τιμές του δείγματός μας

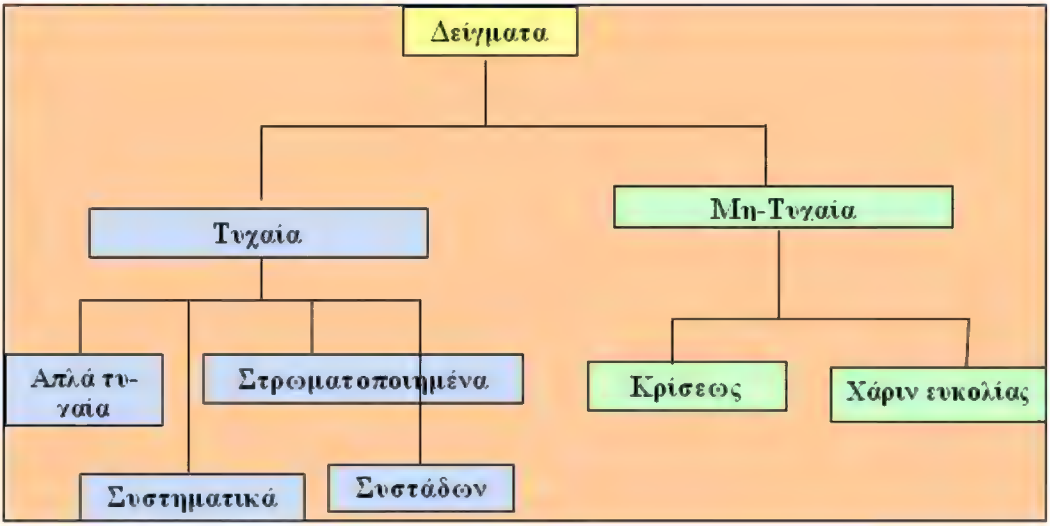
Πίνακας 7.1 Συνάρτηση πιθανοφάνειας διωνυμικής μεταβλητής

p_i	L(1,1,0)
0,0	0,0000
0,1	0,0090
0,2	0,0320
0,3	0,0630
0,4	0,0960
0,5	0,1250
0,6	0,1440
0,7	0,1470

7.2.2 Σχέδια δειγματοληψίας

Διακρίνουμε τα δειγματοληπτικά σχέδια σε δύο κύριες κατηγορίες, τα τυχαία και τα μητυχαία σχέδια. Επαναλαμβάνουμε για να τονιστεί ότι, το χαρακτηριστικό των τυχαίων δειγμάτων είναι ότι μόνο σε αυτά μπορούμε να χρησιμοποιήσουμε την πιθανοθεωρία για να γενικεύσουμε τα συμπεράσματα που τα αφορούν, βάσει των τιμών των στατιστικών του $\bar{X}, Me, Mo, s^2, \beta_1, \beta_2$, των κύριων τεχνικών δειγματοληψίας, δίνεται στο παρακάτω Σχήμα.

Σχήμα 7.1 Δειγματοληπτικά σχέδια



Ο ευκολότερος τρόπος επιλογής απλού τυχαίου δείγματος είναι με τη χρήση του Πίνακα των τυχαίων αριθμών. Τη μήτρα αυτή περιέχουν πολλά λογισμικά, ακόμα και το MS-Excel (random numbers generator). Πρόκειται για πίνακα που περιλαμβάνει άπειρα ψηφία από το σύνολο (0,1,2,...,9). Κατά συνέπεια, η πιθανότητα επιλογής καθενός από αυτά τα ψηφία είναι ίση με οποιουδήποτε άλλου, δηλ. P(0) = P(1) = ... = P(9) = 1/10. Επομένως, εάν ζητείται ένα δείγμα μεγέθους 50 παρατηρήσεων, τότε αναμένεται ότι με πιθανότητα 10% αυτό θα περιλαμβάνει 5 παρατηρήσεις από το ψηφίο 0, 5 παρατηρήσεις από το ψηφίο 1 κοκ. Ο πίνακας των τυχαίων αριθμών αν και δείχνει ταξινομημένος σε στήλες πλάτους 5 ψηφίων η κάθε μία, εντούτοις, αυτό έχει γίνει απλώς και μόνο για να είναι ευκολότερη η ανάγνωση του. Διαφορετικά, η σειρά υποψηφίων (0,1,2,...,9) θα ήταν συνεχής.

Τα βήματα που ακολουθούμε για την επιλογή τυχαίου δείγματος (simple random samle) μέσα από τον πίνακα των τυχαίων αριθμών είναι:

Βήμα 1ο Ορίζουμε το μέγεθος ή πλαίσιο N του πληθυσμού από τον οποίο θέλουμε να αντλήσουμε αντιπροσωπευτικό δείγμα, δηλαδή τυχαίο δείγμα.

Βήμα 2ο Αντιστοιχίζουμε σε καθεμία μονάδα του πληθυσμού έναν αριθμό, έτσι ώστε να έχουμε κωδικοποιήσει τις παρατηρήσεις του.

Βήμα 3ο Με κλειστά μάτια τοποθετούμε το μολύβι στον πίνακα των τυχαίων αριθμών. Το σημείο αυτό είναι το σημείο εκκίνησης.

Βήμα 4ο Από το σημείο εκκίνησης διαβάζουμε αριθμούς με πλήθος στοιχείων όσο είναι το μέγεθος του πληθυσμού. Εάν για παράδειγμα, N=1.000 τότε διαβάζουμε τετραψήφιους αριθμούς προς οποιαδήποτε κατεύθυνση είτε οριζόντια είτε κάθετα. Κάθε τετραψήφιος που έχει μέγεθος μέχρι 1.000 επιλέγεται στο δείγμα. Εάν κατά την εκτέλεση αυτής της διαδικασίας προκύψει να επιλέξουμε δύο φορές τον ίδιο αριθμό, θα αφήσουμε και τους δύο μέσα στο δείγμα εφόσον, όπως και συνιστάται, δειγματοληπτούμε με επανάθεση, δηλ. από άπειρο πληθυσμό. Στην αντίθετη περίπτωση, δειγματοληψίας χωρίς επανάθεση, αριθμός ο οποίος εμφανίζεται δεύτερη φορά η διαγράφεται. Η διαδικασία συνεχίζεται μέχρι να ολοκληρωθεί το πλήθος των στοιχείων του δείγματος, δηλαδή αν το μέγεθος του είναι n=80 θα σταματήσουμε όταν συμπληρώσουμε και την 80η παρατήρηση.

Παράδειγμα 7.3

Για την αξιολόγηση του διδακτικού έργου ακαδημαϊκής μονάδας Ελληνικού Α.Ε.Ι. έχει προσδιοριστεί ότι απαιτείται να μελετηθεί δείγμα 96 φοιτητών. Όταν το σύνολο των εγγεγραμμένων φοιτητών στο τμήμα είναι 2.700. Να δείξετε πώς θα επιλεγεί το υπόψη δείγμα με βάση το σχέδιο της απλής τυχαίας δειγματοληψίας, και μάλιστα όταν αυτή υλοποιείται μέσα από τον πίνακα των τυχαίων αριθμών.

Απάντηση:

Εφαρμόζουμε τη διαδικασία των τεσσάρων βημάτων που εξηγήσαμε παραπάνω. Συγκεκριμένα στην περίπτωση αυτή θα έχουμε:

Βήμα 1ο Εδώ N=2.700 με ζητούμενο μέγεθος δείγματος n=96. Εντούτοις, επειδή είναι γνωστό ότι τουλάχιστον 25% των παρόντων φοιτητών δεν απαντούν, αυξάνουμε ανάλογα το μέγεθος του δείγματος θα είναι n*=96/0,75 = 128.

Βήμα 2ο Η αντιστοίχηση των ονομάτων των φοιτητών με αριθμούς, δηλαδή η κωδικοποίηση του πληθυσμού μας, υπάρχει ήδη από τους καταλόγους της γραμματείας του τμήματος.

Βήμα 3ο Υποθέτουμε ότι τυχαία έχουμε ως σημείο εκκίνησης τον τετραψήφιο 4.928.

Βήμα 4ο Επειδή ο αριθμός εκκίνησης είναι μεγαλύτερος από 2.700 που είναι το όριο του πληθυσμού μας, τον παραλείπουμε και προχωράμε στον επόμενο, με κατεύθυνση οριζόντια, όπως φαίνεται στο παρακάτω σχήμα, και μέχρι τη συμπλήρωση των 128 ζητούμενων αριθμών σύνθεσης του τυχαίου δείγματός μας.

Σχήμα 7.2 Επιλογή τυχαίου δείγματος από τον Πίνακα των τυχαίων αριθμών.

Δείγμα από το μητρώο φοιτητών

α/α	Ονοματεπώνυμο
001	Τιτομαχιλάκη Αργυρή
002	Ζαμπετάκης Κώστας
…	… …
…	… …
…	… …
…	… …
…	… …
2.700	Χονδρόμαλλης Ιωάννης

Τμήμα Πίνακα Τυχαίων Αριθμών

49280	88924	35779	00283	81163	07275
11100	02340	12860	74697	96644	89439
09893	23997	20048	49420	88872	08401

Οι 5 πρώτες δειγματοληπτικές μονάδες (δ.μ.) του δείγματος *n=128 φοιτητών**

δ.μ. # 4.928 – δεν υπάρχει, απορρ. (>2.700)
δ.μ. # 0.889
δ.μ. # 2.435
δ.μ. # 7.790 – δεν υπάρχει, απορρ. (>2.700)
δ.μ. # 0.283
δ.μ. # 8.116 – δεν υπάρχει, απορρ. (>2.700)
δ.μ. # 3.072 – δεν υπάρχει, απορρ. (>2.700)
δ.μ. # 7.511 – δεν υπάρχει, απορρ. (>2.700)
δ.μ. # 1.000
δ.μ. # 2.340

Στη στρωματοποιημένη τυχαία δειγματοληψία (stratified random sampling) ακολουθούμε τα παρακάτω βήματα για τη σύσταση του υπόψη δείγματος:

Βήμα 1ο Χωρίζουμε τον πληθυσμό Ν σε υποσύνολα που εδώ ονομάζονται στρώματα (strata) με κριτήριο την ομοιογένεια κάθε στρώματος, αναφορικά με το χαρακτηριστικό που εκφράζει η τυχαία μας μεταβλητή Χ. Με άλλα λόγια, κάθε στρώμα πρέπει να έχει μικρή μεταβλητότητα εντός του (μικρή σ_X^2), ενώ ταυτόχρονα επιδιώκεται η μεγαλύτερη δυνατή μεταβλητότητα μεταξύ των στρωμάτων. Εάν το μέγεθος του (υπό-) δείγματος σε κάθε στρώμα είναι ανάλογο του μεγέθους του στρώματος στον πληθυσμό, τότε μιλάμε για ανάλογη στρωματοποιημένη δειγματοληψία. Στην αντίθετη περίπτωση, η οποία συμβαίνει όταν η διακύμανση εντός των στρωμάτων δεν είναι η ελάχιστη επιθυμητή, όσο μεγαλύτερη η διακύμανση, τόσο μεγαλύτερο και το μέγεθος του υπο-δείγματος από το υπόψη στρώμα. Με άλλα λόγια, στην τελευταία περίπτωση εφαρμόζουμε δυσανάλογη στρωματοποιημένη δειγματοληψία.

Βήμα 2ο Επιλέγουμε τυχαίο δείγμα εντός κάθε στρώματος, έστω με την προαναφερθείσα απλή διαδικασία του Πίνακα των τυχαίων αριθμών.

Βήμα 3ο Το συνολικό στρωματοποιημένο δείγμα προκύπτει ως το άθροισμα των δειγματοληπτικών μονάδων από το κάθε στρώμα.

Παράδειγμα 7.4

Το τμήμα marketing μεγάλης επιχείρησης επιθυμεί να εκτιμήσει το ποσοστό ανταπόκρισης των εργαζομένων σε σχεδιαζόμενο νέο λογισμικό ενδοεπικοινωνίας. Οι 500 εργαζόμενοι μπορούν να ταξινομηθούν σε δύο μεγάλες κατηγορίες, πτυχιούχους Α.Ε.Ι. και μηπτυχιούχους. Στην κατηγορία των πτυχιούχων ανήκει το 40% των υπαλλήλων ενώ στους μηπτυχιούχους το υπόλοιποι 60%. Ο διευθυντής Marketing, στατιστικολόγος, έχει εκτιμήσει ότι με στρωματοποιημένο τυχαίο δείγμα n=35 ατόμων μπορεί να έχει αξιόπιστη εκτίμηση του υπόψη ποσοστού, αν και από την εμπειρία του γνωρίζει ότι μόνο το 80% των ερωτώμενων απαντούν. Εξηγείστε πώς θα κατασκευαστεί το υπόψη τυχαίο δείγμα.

Απάντηση:

Βήμα 1ο Ο πληθυσμός εδώ είναι N=500. Επειδή το 80% μόνο απαντούν το συνολικό δείγμα πρέπει να αυξηθεί ανάλογα, δηλ. n*=35/0,8≅44. Κατά συνέπεια, επειδή δεν είναι παράλογο να υποθέσουμε ομοιογένεια εντός του καθενός από τα 2 στρώματα, πτυχιούχων και μη, θα πρέπει, από τον Πίνακα των τυχαίων αριθμών, να επιλέξουμε 2 απλά τυχαία δείγματα n_1=44*0,40≅18 πτυχιούχων και n2=44*0,60≅26 μη-πτυχιούχων.

Βήμα 2ο Εκτελούμε για κάθε στρώμα, τα βήματα 2-4 της επιλογής τυχαίου δείγματος από τον πίνακα των τυχαίων αριθμών με την προϋπόθεση τριψήφιων αριθμών αφού ο πληθυσμός N=500. Έτσι συνθέτουμε τις δειγματοληπτικές μονάδες των n_1 και n_2.

Βήμα 3ο Το συνολικό τυχαίο δείγμα της ανάλογης στρωματοποιημένης δειγματοληψίας θα είναι μεγέθους $n = n_1 + n_2$.

Στη συστηματική τυχαία δειγματοληψία (systematic random sampling) για τη σύσταση του τυχαίου δείγματος ακολουθούμε τα εξής δύο απλά βήματα:

Βήμα 1ο Μερίζουμε τον πληθυσμό N σε λ σύνολα, όπου λ=N/n, και n παριστά το μέγεθος του τυχαίου δείγματος που επιθυμούμε να επιλέξουμε με συστηματική τυχαία δειγματοληψία.

Βήμα 2ο Από το πρώτο σύνολο λ_1 επιλέγουμε τυχαία ένα στοιχείο. Έστω ότι αυτό είναι η 4η παρατήρηση (x_1=4η παρατήρηση) και ότι λ=25 με n=20. Συνεχίζουμε με την 4^{η} παρατήρηση στη δεύτερη 25άδα, δηλ. επιλέγουμε ως 2^{η} δειγματοληπτική μονάδα (x_2) την παρατήρηση με κωδικό 29 (=4+25), η 3^{η} τιμή της X (x_3) θα είναι το στοιχείο 54 (=4+25+25), η 4^{η} τιμή της X (x_4) θα είναι το στοιχείο 79 (=4+25+25+25), η 5^{η} τιμή της X (x_5) θα είναι το στοιχείο 104 (=4+25+25+25+25), κοκ, μέχρι την x_{20}.

Η συστηματική τυχαία δειγματοληψία είναι ο ευκολότερος τρόπος επιλογής τυχαίων δειγμάτων από πληθυσμούς ταξινομημένους σε καταλόγους, πχ. τηλεφωνικός κατάλογος, μητρώο υπαλλήλων εταιρείας ή οργανισμού, μητρώο εγγεγραμμένο σε επαγγελματικό ή άλλο επιμελητήριο και γενικά σύνολα για τα οποία είμαστε πεπεισμένοι ότι περιλαμβάνουν όλες τις πληθυσμιακές μονάδες.

Παράδειγμα 7.5

Για το προηγούμενο παράδειγμα να δείξετε πώς θα μπορούσατε να επιλέξετε τυχαίο δείγμα με βάση το συστηματικό σχέδιο.

Απάντηση:

Βήμα 1ο Μερίζουμε τον πληθυσμό των N=500 εργαζομένων σε 44 (=n*=35/0,8≅44) σύνολα των 11 (=λ=N/n=500/44) παρατηρήσεων.

Βήμα 2ο Από την πρώτη 11κάδα $λ_1$ επιλέγουμε τυχαία, έστω, το στοιχείο 7, δηλ. $x_1=7^η$ κωδικοποιημένη (από το μητρώο των υπαλλήλων της εταιρείας) παρατήρηση. Επομένως, η $2^η$ δειγματοληπτική μονάδα (x_2) θα είναι η 18η (=7+11) του μητρώου, η $3^η$ η $29^η$ (=7+11+11), κοκ, μέχρι να συμπληρωθούν οι ζητούμενες 44 παρατηρήσεις του τυχαίου, με το συστηματικό σχέδιο, δείγματος.

Στην τυχαία δειγματοληψία συστάδων (clusters) ο στατιστικός πληθυσμός μερίζεται σε συστάδες ($1^ο$ βήμα), οι οποίες αποτελούν υποσύνολα αντιπροσωπευτικά της δομής του συνολικού πληθυσμού. Στη συνέχεια επιλέγεται τυχαίο δείγμα συστάδων ($2^ο$ στάδιο) οι παρατηρήσεις του οποίου μελετώνται. Ως συστάδες νοούνται «φυσικά ή γεωγραφικά» ενδεχόμενα π.χ. χώρες, εκλογικές περιφέρειες, νομοί, δήμοι πόλεων, γεωγραφικά ορισμένα τμήματα αγορών προϊόντων, κ.λπ.

Από τον ορισμό της τεχνικής της δειγματοληψίας συστάδων αβίαστα προκύπτει ότι, αφενός, εντός κάθε συστάδας πρέπει να υπάρχει μεγάλη ανομοιογένεια, δηλ. υψηλή διασπορά, και αφετέρου, μεταξύ των συστάδων μικρή μεταβλητότητα, δηλ. υψηλός δείκτης ομοιογένειας. Με άλλα λόγια, τα χαρακτηριστικά γνωρίσματα του σχεδίου της τυχαίας δειγματοληψίας συστάδων είναι τα ακριβώς αντίθετα εκείνου της στρωματοποιημένης.

Εντούτοις, η τυχαία δειγματοληψία συστάδων, αν και γενικά απαιτεί μεγαλύτερα μεγέθη δειγμάτων, για να δώσει αξιόπιστα αποτελέσματα, κοστίζει όμως λιγότερο σε χρόνο και χρήμα, ειδικά από το απλό τυχαίο δείγμα, εάν μάλιστα πρόκειται για ανθρώπινους πληθυσμούς πολύ απομακρυσμένους γεωγραφικά.

Από τα μη-τυχαία **δειγματοληπτικά σχέδια** (non-probability samples) συνηθέστερα χρησιμοποιούμενα είναι το «χάριν ευκολίας» (convenience sample), και το «κρίσεως» (judgment sample).

Στη «χάριν ευκολίας» τεχνική, οι δειγματοληπτικές μονάδες συλλέγονται χωρίς καμία φροντίδα τυχαιότητας, αντίθετα μάλιστα, η επιλογή τους γίνεται με γνώμονα την ευκολία ή το χαμηλό κόστος ανακάλυψής τους. Για παράδειγμα δείγμα ευκολίας συνήθως χρησιμοποιούν οι δημοσιογραφικές έρευνες τηλεοπτικών σταθμών, με ορισμένα ερωτήματα πολιτικής ή οικονομικής φύσεως. Τέτοιο δείγμα ευκολίας μπορεί να είναι οι 10 πρώτοι περαστικοί πολυσύχναστης διασταύρωσης στην πρωτεύουσα. Σε ορισμένες μάλιστα περιπτώσεις οι δειγματοληπτικές μονάδες στα δείγματα ευκολίας μπορεί να είναι «αυτο-επιλεγόμενες»! Για παράδειγμα, πολλές εταιρείες στις ιστοσελίδες τους ζητούν τη γνώμη των επισκεπτών τους, οι οποίοι υποβάλλουν τα συμπληρωμένα ερωτηματολόγια ηλεκτρονικά κατά την ώρα επίσκεψης τους στις υπόψη ιστοσελίδες.

Στην τεχνική της δειγματοληψίας κρίσεως, συντίθεται μη-τυχαίο δείγμα με κριτήριο τη συνάφεια-ειδικότητα των ερωτώμενων στο συγκεκριμένο θέμα. Σε ορισμένες περιπτώσεις δεν υπάρχει άλλη δυνατότητα επιλογής δειγματοληπτικών μονάδων εκτός από το σχέδιο κρίσεως. Για παράδειγμα, όταν δημοσιογράφος ερευνά π.χ. ιατρικό θέμα στην εκπομπή του, αναπόδραστα, ψάχνει για συνέντευξη με κάποιους διεθνούς φήμης στο θέμα αυτό γιατρούς. Ή ακόμα, όταν διευθυντής αίθουσας διαπραγμάτευσης αξιών, πρόκειται να αποφασίσει για αγορά 5.000.000$ στα επόμενα 5 λεπτά, απλά θα χρησιμοποιήσει την «αίσθηση» που έχουν την ώρα εκείνη οι dealers στην αίθουσα, στη βάση των βραχυπρόθεσμων αναλύσεων της τάσης της αρμόδιας διεύθυνσης.

Προφανώς, αφού δεν μεθοδεύεται τυχαιότητα (δηλ. ίση και γνωστή εκ των προτέρων πιθανότητα επιλογής, αφενός, κάθε μονάδας στο δείγμα, και αφετέρου, κάθε ισομεγέθους δείγματος έναντι των υπολοίπων), στη λήψη των παρατηρήσεων των δύο τελευταίων μη τυχαίων σχεδίων, δεν μπορεί να χρησιμοποιηθεί η πιθανοθεωρία για να γενικεύσουμε τα συμπεράσματα που προκύπτουν από τις τιμές των στατιστικών των σχετικών δειγμάτων. Επομένως, με τα δείγματα αυτά μόνο περιγραφική ανάλυση μπορούμε να κάνουμε. Δεν μπορούμε να προχωρήσουμε σε στατιστική συμπερασματολογία ή επαγωγή από το δείγμα στον πληθυσμό.

7.2.3 Σφάλματα δειγματοληπτικών ερευνών

- **Σφάλμα κάλυψης ή μεροληψία επιλογής.**

Εμφανίζεται όταν κάποιες ομάδες αποκλείονται από τον ερευνώμενο πληθυσμό και δεν έχουν πιθανότητα να επιλεγούν.

- **Σφάλμα ή μεροληψία μη-απάντησης.**

Πολίτες που δεν απαντούν μπορεί να διαφέρουν από εκείνους που απαντούν στην έρευνα.

- **Δειγματοληπτικό σφάλμα**

Πάντα θα υπάρχει μεταβλητικότητα από δείγμα σε δείγμα, με την έννοια ότι συμπεράσματα από ένα δείγμα, ακόμα και τυχαίο, δεν θα συμπίπτουν ποτέ με εκείνα άλλου τυχαίου δείγματος. Πρόκειται για τις λεγόμενες τυχαίες κυμάνσεις της δειγματοληψίας.

- **Σφάλμα μέτρησης**

Μη-ακριβής διατύπωση της ερώτησης από το δειγματολήπτη οδηγεί τον ερωτώμενο να απαντήσει για κάτι άλλο από αυτό που ζητά ο σχεδιαστής του ερωτηματολογίου.

7.3 Κατανομές δειγματοληψίας μονομεταβλητών πληθυσμών

7.3.1 Βασικές έννοιες κατανομών δειγματοληψίας

Η στατιστική συμπερασματολογία επιτυγχάνεται κυρίως με δύο εναλλακτικές μεθοδολογίες. Στην πρώτη, Κλασική επαγωγική στατιστική, τα συμπεράσματα για τις άγνωστες πληθυσμιακές παραμέτρους (π.χ. μέσος, διακύμανση, αναλογία κ.λπ.) εκτιμώνται από τις τιμές των αντίστοιχων στατιστικών συναρτήσεων των τιμών του τυχαίου δείγματος. Στη δεύτερη, Μπαγιεζιανή (Bayesian statistical inference) επαγωγική συμπερασματική, οι πληροφορίες από το τυχαίο δείγμα συνδυάζονται με διαθέσιμη εκ των προτέρων (a priori) εξωγενή πληροφόρηση. Στο τεύχος αυτό δεν ασχολούμαστε με Μπαγιεζιανή στατιστική.

Η Κλασική Επαγωγική συμπερασματολογία που παρουσιάζουμε εδώ, περιλαμβάνει δύο στάδια προσέγγισης. Το πρώτο αναφέρεται στην Εκτίμηση των πληθυσμιακών παραμέτρων και το δεύτερο συνίσταται στον (παραμετρικό) Έλεγχο των Στατιστικών Υποθέσεων. Γι' αυτά τα δύο στάδια θα μιλήσουμε στο επόμενο κεφάλαιο.

'Όταν λέμε εκτίμηση (estimate) μιας (πληθυσμιακής) παραμέτρου, έστω π.χ. του μέσου μ του πληθυσμού της X τυχαίας μεταβλητής, από τα δεδομένα του τυχαίου δείγματος με τιμές x_i, (i=1,2,...,n), εννοούμε μια προσεγγιστική τιμή της άγνωστης πληθυσμιακής που μας δίνει η αντίστοιχη συνάρτηση του δείγματος. Για το μέσο αριθμητικό π.χ., η συνάρτηση ως γνωστόν έχει τη μορφή:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$	**(7.3)**

Κατά συνέπεια πρέπει να είναι πλέον απολύτως σαφές στον αναγνώστη ότι, η εκτίμηση του άγνωστου αλλά αληθινού πληθυσμιακού μέσου μ γίνεται από την τιμή του δειγματικού μέσου $\bar{X}$ (σχέση 7.3). Αυτή είναι στατιστική συνάρτηση των τυχαίων μεταβλητών-παρατηρήσεων X_i, (i=1,2,...,n) του τυχαίου δείγματος. Έτσι, για λόγους διάκρισης ο δειγματικός μέσος ($\bar{X}$) ονομάζεται στατιστική ή εκτιμήτρια (estimator) του άγνωστου πληθυσμιακού μ όπου αντίστοιχα εάν κάναμε απογραφή {X_i, (i=1,2,...,N)} θα υπολογιζόταν από τη σχέση:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$	**(7.4)**

Βέβαια ως στατιστική θα μπορούσαμε να ορίσουμε οποιαδήποτε συνάρτηση των παρατηρήσεων του δείγματος. Στο βιβλίο αυτό, όμως θα εννοούμε τις δειγματικές συναρτήσεις των μέσου ($\bar{X}$), διακύμανσης (s^2), τυπικής απόκλισης (s), αναλογίας (p) κ.λπ. Επισημαίνεται ότι, έχει καθιερωθεί στη διεθνή βιβλιογραφία να συμβολίζονται με Λατινικούς χαρακτήρες οι στατιστικές του δείγματος και με Ελληνικούς οι άγνωστες παράμετροι του πληθυσμού. Οπότε, για τις προαναφερθείσες στατιστικές, συμβολίζουμε τις παραμέτρους τους, αντίστοιχα, ως: μ, σ^2, σ, P.

Πρέπει να τονιστεί ότι, οι στατιστικές, ως συναρτήσεις των παρατηρήσεων {x_i, (i=1,2,...,n)} του τυχαίου δείγματος, οι οποίες, όπως έχουμε ήδη εξηγήσει, είναι τυχαίες μεταβλητές, είναι και αυτές (οι στατιστικές ή εκτιμήτριες) τυχαίες μεταβλητές.

Επομένως, οι στατιστικές ως τυχαίες μεταβλητές έχουν και κατανομές πιθανότητας, μέσω των οποίων ορίζονται πιθανοθεωρητικά. Οι τελευταίες ονομάζονται Κατανομές Δειγματοληψίας οποιασδήποτε στατιστικής.

Πιο συγκεκριμένα, ορίζουμε ως Κατανομή Δειγματοληψίας της Στατιστικής Θ, την κατανομή συχνότητας των τιμών της (Θ_i), οι οποίες προκύπτουν, εάν από πληθυσμό (X) μεγέθους (N) με τυχαία διαδικασία φτιάξουμε όλα τα δυνατά και ισομεγέθη δείγματα, μεγέθους (n), και για κάθε ένα από αυτά υπολογίσουμε και καταγράψουμε την τιμή της στατιστικής Θ.

Παράδειγμα 7.6

Με βάση την κατανομή πιθανότητας της τυχαίας μεταβλητής X στο Παράδειγμα 7.1, που επαναλαμβάνουμε εδώ για ευκολία, να κατασκευαστούν οι κατανομές δειγματοληψίας των στατιστικών του μέσου ($\bar{X}$) και της διακύμανσης (s^2), υποθέτοντας ότι φτιάχνετε δείγματα μεγέθους n=2:

$$f(x)=\begin{cases} 1/4 & \text{για} \quad x=0 \\ 2/4 & \text{για} \quad x=1 \\ 1/4 & \text{για} \quad x=2 \\ 0 & \text{για} \quad x\neq 0,1,2 \end{cases}$$

Απάντηση:

Οι τιμές xi δεν έχουν όλες την ίδια πιθανότητα, f(X=0)=f(X=2)=1/4 ενώ f(X=1)=2/4. Αφού ζητείται να κατασκευαστούν όλα τα δυνατά δείγματα μεγέθους n=2 , εννοείται ότι το σύνολο του δειγματικού αυτού χώρου (τυχαία δειγματοληψία με επανάθεση) θα είναι 3^2=9 δείγματα. Για κάθε ένα από τα δείγματα υπολογίζουμε το μέσο ($\bar{X}$) και τη διακύμανσή του (s^2). Στη συνέχεια ταξινομούμε σε κατανομή σχετικής συχνότητας τις τιμές των δύο στατιστικών που μόλις βρήκαμε. Οι δύο αυτές κατανομές συνιστούν τις ζητούμενες κατανομές δειγματοληψίας.

Οι υπολογισμοί αυτοί δίνονται στον παρακάτω Πίνακα.

Πίνακας 7.2 Κατασκευή κατανομών δειγματοληψίας μέσου και διακύμανσης

Κατανομή δείγματος		Υπολογισμοί για τις τιμές $\bar{X}_i$ και s_i^2		Κατανομή δειγματοληψίας $\bar{X}$		Κατανομή δειγματοληψίας s^2	
(x_1, x_2)	$f(x_1,x_2) = f(x_1)f(x_2)$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2$	$\bar{X}_i$	$f(\bar{X}_i)$	s_i^2	$f(s_i^2)$
(0, 1)	0,1250=2/16 =(1/4)(2/4)	0,5 =(0+1)/2	0,50 $=(0-0,5)^2+(1-0,5)^2$	0,0	1/9 =0,1111	0,0	3/9 =0,3333
(0, 2)	0,0625	1,0	2,00	0,5	2/9= 0,2222	0,5	4/9 =0,4444
(1, 2)	0,1250	1,5	0,50	1,0	3/9 =0,3333	2,0	2/9 =0,2222
(1, 0)	0,1250	0,5	0,50	1,5	2/9 =0,2222		
(2, 0)	0,0625	1,0	2,00	2,0	1/9 =0,1111		
(2, 1)	0,1250	1,5	0,50				
(0, 0)	0,0625	0,0	0,00				
(1, 1)	0,2500	1,0	0,00				
(2, 2)	0,0625	2,0	0,00				
	1,0000				1,0000		1,0000

Ας σημειωθεί ότι μόνο για εκπαιδευτικούς σκοπούς παρουσιάζεται ο τρόπος κατασκευής των κατανομών δειγματοληψίας του μέσου και της διακύμανσης τυχαίας μεταβλητής, πολύ μικρού πληθυσμού. Στην πράξη ποτέ δεν κατασκευάζουμε την κατανομή δειγματοληψίας κάποιας στατιστικής πριν να κάνουμε Επαγωγική Συμπερασματική γι' αυτήν. Απλώς εφαρμόζουμε τα αποτελέσματα της έρευνας των στατιστικολόγων οι οποίοι στην προσπάθεια τους να «υποδειγματοποιήσουνε» τις διαδικασίες έχουν ανακαλύψει τα θεωρητικά πρότυπα που αυτές (οι στατιστικές ή εκτιμήτριες) ακολουθούν. Έτσι, η επαγωγή γίνεται χρησιμοποιώντας τις γνωστές αυτές θεωρητικές κατανομές πιθανότητας όπως π.χ. την τυπική κανονική (Ζ) που είδαμε στο προηγούμενο κεφάλαιο, αλλά και τις χ^2, t-student και F (Fisher) τις οποίες συνοπτικά θα παρουσιάσουμε στον παρόν κεφάλαιο.

Παρακάτω θα αναφερθούμε ενδεικτικά σε τρεις μόνο κατανομές δειγματοληψίας η πιθανοθεωρητική συμπεριφορά των οποίων περιγράφεται ικανοποιητικά από γνωστούς θεωρητικούς νόμους.

7.3.2 Κατανομή δειγματοληψίας του μέσου

Όπως ήδη πρέπει να έχει γίνει σαφές, κατανομή δειγματοληψίας του μέσου ($\bar{X}$) είναι η κατανομή των σχετικών συχνοτήτων των μέσων ($\bar{X}$, i=1,2,...,k) k ισομεγεθών δειγμάτων μεγέθους (n), τα οποία ελήφθησαν με τυχαία δειγματοληψία, από πληθυσμό μεγέθους N της τυχαίας μεταβλητής X.

Η μορφή αυτής της κατανομής, δηλαδή η καμπύλη πιθανοτήτων της (στο όριο οι σχετικές συχνότητες ορίζουν εμπειρικές πιθανότητες), εξαρτάται από τρεις παράγοντες:

α) την κατανομή πιθανότητας του γεννήτορα πληθυσμού της, π.χ. αν $X \sim N(\mu,\sigma^2)$,

β) εάν είναι γνωστή ή άγνωστη η πληθυσμιακή διακύμανση (σ^2) και

γ) το μέγεθος (n) των ισομεγεθών δειγμάτων, μικρό ($n<30$) ή μεγάλο ($n>30$).

Οι γνωστές θεωρητικές κατανομές πιθανότητας που ακολουθεί η στατιστική του μέσου ($\bar{X}$), ανάλογα με το συνδυασμό των τριών παραπάνω παραγόντων, δίνονται στον παρακάτω συνοπτικό Πίνακα.

Πίνακας 7.3 Κατανομές δειγματοληψίας του μέσου ($\bar{X}$)

σ^2 / X~	Γνωστή	Άγνωστη $n \geq 30$	Άγνωστη $n < 30$
Οποιοδήποτε	$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$	Κ.Ο.Θ.	Απαραμετρική Στατιστική
Άπειρος Κανονικός	$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$	$\bar{X} \sim N(\mu, \frac{s^2}{n})$	$\bar{X} \sim t_{(n-1)}$
Πεπερασμένος Κανονικός	$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}\frac{N-n}{N-1})$	--	--

Όπου
- (N-n)/(N-1) είναι ο διορθωτικός παράγοντας της πληθυσμιακής διακύμανσης στην περίπτωση πεπερασμένου πληθυσμού ή δειγματοληψίας χωρίς επανάθεση.

- Κ.Ο.Θ. Κεντρικό Οριακό Θεώρημα: πάρα πολύ σπουδαίο θεώρημα που απέδειξαν οι στατιστικολόγοι με τεράστια πρακτική αξία αφού λέει με απλά λόγια ότι «εάν δειγματοληπτούμε από οποιοδήποτε πληθυσμό που έχει μέσο μ και διακύμανση σ2 τότε όσο το μέγεθος του τυχαίου δείγματος αυξάνει τόσο η κατανομή δειγματοληψίας του μέσου ($\bar{X}$) προσεγγίζει την κανονική με μέσο μ και διακύμανση σ^2/n». Συνοπτικά γράφουμε για το κεντρικό οριακό θεώρημα:

$$\bar{X}_{n \to \infty} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \textbf{(7.5)}$$

Από τη σχέση (7.5) αβίαστα προκύπτει ότι η παρακάτω οριζόμενη τυπική μεταβλητή Z* ακολουθεί την τυπική κανονική, τις πιθανότητες της οποίας γνωρίζουμε πλήρως αφού την έχουν πινακοποιήσει οι Gauss-Laplace και επομένως μπορούμε εύκολα να κάνουμε εκτιμήσεις.

$Z^* \equiv \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \sim Z(0,1)$	**(7.6)**

Κάτω από τις προϋποθέσεις του κεντρικού οριακού θεωρήματος.

Εντούτοις, ας σημειωθεί ότι όταν η πληθυσμιακή διακύμανση είναι άγνωστη και δειγματοληπτούμε από «μικρό», δηλ. πεπερασμένο πληθυσμό, τότε η κατανομή δειγματοληψίας του μέσου είναι άγνωστη δηλ. δεν μπορούμε να κάνουμε εκτιμήσεις για την άγνωστη πληθυσμιακή παράμετρο μ με βάση την τιμή της εκτιμήτριας ($\bar{X}$), της διακύμανσης (σ^2 ή s^2) και το μικρό ($n<30$) ή μεγάλο ($n>30$) μέγεθος του τυχαίου δείγματος.
Με βάση λοιπόν τις γνωστές περιπτώσεις κατανομών δειγματοληψίας, οι θεωρητικές κατανομές πιθανότητας που χρησιμοποιούμε για τη στατιστική συμπερασματολογία (κλασική παραμετρική στατιστική επαγωγή) για τον πληθυσμιακό μέσο δίνονται στον παρακάτω εξίσου πολύ σημαντικό Πίνακα.

Πίνακας 7.4 Τυπική κανονική και t-student οι θεωρητικές κατανομές για τις γνωστές στατιστικές των κατανομών δειγματοληψίας του μέσου ($\bar{X}$)

σ^2 / Κατ.Πληθ.	Γνωστή	Αγνωστη $n \geq 30$	Αγνωστη $n < 30$
Οποιοσδήποτε	$z^* \equiv \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Z(0,1)$	$z^* \equiv \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim Z(0,1)$	Απαραμετρική Στατ
Απειρος Κανονικός	$z^* \equiv \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Z(0,1)$	$z^* \equiv \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim Z(0,1)$	$t^* \equiv \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$
Πεπερασμένος Κανονικός	$z^* \equiv \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}}} \sim Z(0,1)$	-	-

Παράδειγμα 7.7

Ορκωτός Ελεγκτής για τον έλεγχο των 1.000 εισπρακτέων λογαριασμών εταιρείας με μ=298€ και σ=54€ επιλέγει τυχαίο δείγμα n=36 τιμολογίων. Ζητούνται οι πιθανότητες: **α)** P($\bar{X}$<280) **β)** P($\bar{X}$>320) και **γ)** P(274≤$\bar{X}$≤320).

Απάντηση:

Αφού είναι γνωστή η πληθυσμιακή διακύμανση (s^2) γνωρίζουμε ότι η εκτιμήτρια ($\bar{X}$) του μέσου ακολουθεί την κανονική, ή $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Άρα,

$E(\bar{X}) = \mu = 298$ ενώ το τυπικό σφάλμα του μέσου θα είναι: .

όπου $\sigma_{\bar{x}}$ είναι η τυπική απόκλιση της κατανομής δειγματοληψίας του μέσου που ονομάζεται τυπικό σφάλμα του μέσου. Το τυπικό σφάλμα θα γραφόταν $s_{\bar{x}}$ εάν η διακύμανση του πληθυσμού ήταν άγνωστη και χρησιμοποιούσαμε στη θέση τ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{54}{\sqrt{36}} = 9$

εκτιμήτριά της, τη δειγματική τυπική απόκλιση:

$$s_{\bar{x}} = \sqrt{\frac{1}{n-1}\sum_i (\bar{X}_i - \bar{\bar{X}})^2}$$

α)

$$P(\bar{X} < 280 \mid \mu = 298, \sigma_{\bar{x}} = 9) = P(\frac{\bar{X}-\mu}{\sigma_{\bar{x}}} < \frac{280-298}{9}) = P(z < -2) = 0,228$$

δηλ. υπάρχει 22,8% πιθανότητα η σημειακή εκτ... μέσου μ όπως αυτή εκφράζεται από το δειγματικό $\bar{X}$ να δίνει μεση αξια εισπρακτεων μικρότερη από 28

β)

$$= P(z > 2.44) = P(z < -2.44) = 0,0073$$

δηλ. υπάρχει 0,73% πιθανότητα η σημειακή εκτίμηση του πληθυσμιακού μέσου μ όπως αυτή εκφράζεται από το δειγματικό $\bar{X}$ να δίνει μέση αξία εισπρακτέων μεγαλύτερη από 320€

γ) $P(274 \le \bar{X} \le 320 \mid \mu = 298, \sigma_{\bar{x}} = 9) = P(\frac{\bar{X}_1 - \mu}{\sigma_{\bar{x}}} \le \bar{X} \le \frac{\bar{X}_2 - \mu}{\sigma_{\bar{x}}}) = P(z_1 \le Z \le z_2)$

$$z_1 = \frac{\bar{X}_1 - \mu}{\sigma_{\bar{x}}} = \frac{274-298}{9} = -2,66 \qquad z_2 = \frac{\bar{X}_2 - \mu}{\sigma_{\bar{x}}} = \frac{320-298}{9} = 2,44$$

δηλ. υπάρχει 98,9% πιθανότητα η σημειακή εκτίμηση του πληθυσμιακού μέσου μ όπως αυτή εκφράζεται από το δειγματικό $\bar{X}$ να δίνει μέση αξία εισπρακτέων στο διάστημα (274, 320)€.

7.3.3 Αναγκαία παρένθεση. Δειγματοληψία από κανονικούς πληθυσμούς: Βασικές έννοιες των κατανομών χ^2, t-Student, και F

Έστω X(i), i=1,2,...,ν ανεξάρτητες τυχαίες μεταβλητές, οι οποίες όλες ακολουθούν την τυπική κανονική, δηλ. X(i) ~ N(0,1). Εάν πάρουμε το άθροισμα των τετραγώνων τους,

τότε η νέα μεταβλητή, έστω $X=X(1)^2+X(2)^2+.....+X(v)^2$, δεν ακολουθεί την τυπική κανονική αλλά τη λεγόμενη χ^2 ή νόμο του Pearson κατανομή πιθανότητας, η οποία είναι ειδική περίπτωση της Γάμμα (Γ) συνάρτησης πυκνότητας πιθανότητας και έχει την εξής μορφή:

$f(x)=\begin{cases} =\dfrac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\frac{\nu-2}{2}}e^{-\frac{x}{2}}, & \gamma\iota\alpha \quad x>0 \\ =0 & ,\ \gamma\iota\alpha \quad x\leq 0 \end{cases}$	**(7.7)**

E(X)=v και Var(X)=2v, ενώ η μορφή της χ^2 εξαρτάται από τους βαθμούς ελευθερίας v. Οι βαθμοί ελευθερίας εκφράζουν τις πραγματικά ανεξάρτητες παρατηρήσεις για τις οποίες υπολογίστηκε η υπόψη συνάρτηση πιθανότητας. Εάν για παράδειγμα υπάρχει έστω μία, ακόμα και γραμμική, σχέση μεταξύ τους, τότε λέμε ότι έχουμε v-1 βαθμούς ανεξάρτητων τυχαίων μεταβλητών-παρατηρήσεων στη διάθεσή μας. Έτσι, π.χ. στη συνάρτηση της διακύμανσης (s_X^2) χρησιμοποιείται η τιμή του μέσου ($\bar{X}$) που είναι γραμμική συνάρτηση των τιμών της τυχαίας μεταβλητής X. Επομένως στη δειγματική διακύμανση, έχουμε v-1 βαθμούς ελευθερίας, και ακριβώς γι' αυτόν τον λόγο για να είναι αμερόληπτη εκτιμήτρια (s^2), της άγνωστης πληθυσμιακής (σ^2) υπολογίζεται με παρονομαστή (n-1).

Η συνάρτηση πιθανότητας της χ^2 για ορισμένους βαθμούς ελευθερίας έχει πινακοποιηθεί και είναι διαθέσιμη είτε στα στατιστικά εγχειρίδια είτε στα σχετικά λογισμικά. Στα βιβλία στατιστικής οι πίνακες της χ^2 στο κύριο σώμα τους δίνουν τις κρίσιμες τιμές χ^2 πάνω από τις οποίες υπάρχει η πιθανότητα που δίνεται στην πρώτη γραμμή του Πίνακα για τους δεδομένους βαθμούς ελευθερίας που αναγράφονται στην πρώτη στήλη. Επίσης, ακόμα και το MS-Excel έχει δύο συναρτήσεις, αφενός, για τον υπολογισμό των πιθανοτήτων κάτω από την καμπύλη χ^2, με δεδομένους τους βαθμούς ελευθερίας και την τιμή της χ^2 πάνω από την οποία υπάρχει η υπόψη πιθανότητα εμφάνισης, και αφετέρου, το αντίθετο, δηλ. για δεδομένους βαθμούς ελευθερίας οι οποίοι ορίζουν συγκεκριμένη κάθε φορά καμπύλη, και αθροιστική πιθανότητα δίνει την κρίσιμη τιμή χ^2.

Στο παρακάτω παράδειγμα καλείται ο αναγνώστης να επιβεβαιώσει τις πιθανότητες ή αντίστροφα τις κρίσιμες τιμές χ^2 για την εμπέδωση της χρήσης του σχετικού πίνακα.

Παράδειγμα 7.8

Να υπολογίσετε τις πιθανότητες $f(\chi^2_\nu)$, για ν βαθμούς ελευθερίας:

α) $P(\chi^2_8 > 17,53)$, β) $P(\chi^2_{10} > 18,31)$, γ) $P(\chi^2_{15} > 14,34)$,

δ) $P(\chi^2_{30} > 43,77)$, ε) $P(\chi^2_{50} > 32,36)$.

Επίσης να βρείτε τις ανώτατες τιμές χ^2_ν για τις οποίες ισχύουν οι παρακάτω πιθανότητες:

στ) $P(\chi^2_{50} > \chi^2_{50_0,01}) = 0,01$, ζ) $P(\chi^2_{26} > \chi^2_{26_0,05}) = 0,05$, η) $P(\chi^2_{20} > \chi^2_{20_0,10}) = 0,10$,

θ) $P(\chi^2_{10} > \chi^2_{10_0,975}) = 0,975$, ι) $P(\chi^2_{15} > \chi^2_{15_0,95}) = 0,95$..

Απάντηση:

α) $P(\chi^2_8 > 17,53) = 0,025$, β) $P(\chi^2_{10} > 18,31) = 0,05$, γ) $P(\chi^2_{15} > 14,34) = 0,50$,

δ) $P(\chi^2_{30} > 43,77) = 0,05$, ε) $P(\chi^2_{50} > 32,36) = 0,975$

στ) $P(\chi^2_{50} > 76,15) = 0,01$, ζ) $P(\chi^2_{26} > 38,89) = 0,05$,

η) $P(\chi^2_{20} > 28,41) = 0,10$, θ) $P(\chi^2_{10} > 3,25) = 0,975$,

ι) $P(\chi^2_{15} > 7,26) = 0,95$.

Αν δύο τυχαίες μεταβλητές X και Y έχουν η πρώτη την τυποποιημένη κανονική συνάρτηση πιθανότητας και η δεύτερη την χ^2_ν, δηλ. X~N(0,1) και Y~χ^2_ν ' τότε η παρακάτω μεταβλητή που προκύπτει από αυτές ακολουθεί την t-Student, γνωστή κατανομή πιθανότητας, με ν βαθμούς ελευθερίας. Με άλλα λόγια:

$t \equiv \frac{X}{\sqrt{Y/\nu}} \sim t_\nu$, $X(\equiv \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}) \sim N(0,1)$ και $Y(\equiv \frac{(n-1)\cdot s^2}{\sigma^2}) \sim \chi^2_\nu$	**(7.8)**

Η κατανομή αυτή ανακαλύφθηκε από τον Άγγλο στατιστικό W.S.Gosset το 1908 και έχει πολύ μεγάλη χρησιμότητα σε πολλές περιπτώσεις όπου δειγματοληπτούμε από κανονικό ή καθ' υπόθεση κανονικό πληθυσμό αλλά δεν έχουμε την «πολυτέλεια» χρήσης μεγάλων δειγμάτων, δηλ. με n<30.

Η μορφή της t-student είναι συμμετρική ως προς την τιμή t=0 και το γράφημά της προσδιορίζεται πλήρως μόνο από την παράμετρο των βαθμών ελευθερίας ν.

Χαρακτηριστικό της t-student είναι ότι αυτή προσεγγίζει την τυποποιημένη κανονική [N(0,1)] όσο το μέγεθος του δείγματος αυξάνει, και συγκεκριμένα για ν(=n-1)≥30.

Όπως και στην περίπτωση της συνάρτησης χ^2_ν, έτσι και εδώ στην t-student, για ορισμένους βαθμούς ελευθερίας (t_ν) αυτή έχει πινακοποιηθεί και είναι διαθέσιμη είτε στα στατιστικά εγχειρίδια είτε στα σχετικά λογισμικά. Στα βιβλία στατιστικής οι πίνακες της (t_ν) στο κύριο σώμα τους συνήθως δίνουν τις κρίσιμες τιμές t_{ν_α} όπου α δίνεται από την πιθανότητα $P(|t_7|>1,895)$, δηλ. πάνω από τις οποίες στις ουρές της συμμετρικής αυτής κατανομής υπάρχει η πιθανότητα α/2. Η αθροιστική αυτή πιθανότητα δίνεται στην πρώτη γραμμή του Πίνακα, ενώ στην πρώτη στήλη αναγράφονται οι βαθμοί ελευθερίας. Προφανώς κάθε γραμμή αντιστοιχεί σε διαφορετική καμπύλη t_ν.

Επίσης, ακόμα και το MS-Excel έχει δύο συναρτήσεις, αφενός, για τον υπολογισμό των πιθανοτήτων κάτω από την καμπύλη t_ν, και αφετέρου, το αντίθετο, δηλ. για δεδομένους βαθμούς ελευθερίας οι οποίοι ορίζουν συγκεκριμένη κάθε φορά καμπύλη, και αθροιστική πιθανότητα α δίνει την κρίσιμη τιμή t_{ν_α}.

Παράδειγμα 7.9

Να υπολογίσετε τις πιθανότητες f(t_ν), για ν βαθμούς ελευθερίας:

α) $P(|t_7|>1,895)$, β) $P(t_{11}>0,697)$, γ) $P(t_{26}<2,056)$. .

Επίσης να βρείτε τις κριτικές τιμές t_{ν_α} για τις οποίες ισχύουν οι παρακάτω πιθανότητες:

δ) $P(|t_{10}|>t_{10_0,10})=0,10$, ε) $P(t_{15}>t_{15_0,05})=0,05$, στ) $P(t_{25}<t_{25_0,01})=0,01$.

Απάντηση:

α) $P(|t_7|>1,895)=0,10$, δηλ. υπάρχει πιθανότητα συνολικά 10% στις 2 ουρές της t_7 στο δικατάληκτο όριο |1,895| και κάτω από την καμπύλη $f(t_7)$.

Εναλλακτικά μπορούμε να γράψουμε:

$P(-1,895 < t_7 < 1,895) = 0,90$ ή $P(t_7 < -1,895$ και $t_7 > 1,895) = 0,10$

β) $P(t_{11} > 0,697) = 0,25$,

γ) $P(t_{26} < 2,056) = 0,025$.

δ) $t_{10_0,10} = 1,812$: $P(|t_{10}| > 1,812) = 0,10$,

ε) $t_{15_0,05} = 1,753$: $P(t_{15} > 1,753) = 0,05$,

στ) $t_{25_0,01} = -2,485$: $P(t_{25} < -2,485) = 0,01$, δηλ. η μονοκατάληκτη κάτω κριτική τιμή της $t_{25_0,01}$, για να αφήνει κάτω από αυτήν μόνο το 1% του εμβαδού στην αριστερά ουρά υπό την t_{25} πρέπει να είναι η -2,485.

Επίσης πολύ χρήσιμη κατανομή πιθανότητας στις εφαρμογές και ιδιαίτερα στην ανάλυση διακύμανσης, είναι η κατανομή F που ορίζεται ως ο λόγος δύο άλλων κατανομών χ^2_v. Πιο συγκεκριμένα αποδεικνύεται το ακόλουθο θεώρημα:

Αν Y(1) και Y(2) δύο ανεξάρτητες τυχαίες μεταβλητές με v_1 και v_2 αντίστοιχα, βαθμούς ελευθερίας, τότε η τυχαία μεταβλητή που σχηματίζεται ως λόγος τους, διαιρεμένος με τους αντίστοιχους βαθμούς ελευθερίας σε κάθε μία, ακολουθεί την F με v_1 και v_2 ή $F_{(v_1,v_2)}$ δηλ.

$F \equiv \dfrac{X(1)/v_1}{X(2)/v_2} \sim F_{(v_1,v_2)}$, $X(1) \sim \chi^2_{v_1}$ και $X(2) \sim \chi^2_{v_2}$	**(7.9)**

Όπως και στην περίπτωση των δύο παραπάνω συναρτήσεων χ^2_v και t_v, έτσι και εδώ στην $F_{(v_1,v_2)}$, για ορισμένους βαθμούς ελευθερίας αριθμητή και παρονομαστή (v_1, v^2) αυτή έχει πινακοποιηθεί και είναι διαθέσιμη, συνήθως για ανώτατες κρίσιμες τιμές σε α=5% ή α=1%, είτε στα στατιστικά εγχειρίδια είτε στα σχετικά λογισμικά.

Στα βιβλία στατιστικής οι πίνακες της $F_{(v_1,v_2)}$ στο κύριο σώμα τους συνήθως δίνουν τις ανώτατεςκρίσιμεςτιμές $F_{(v_1,v_2),\alpha}$, όπουαδίνεταιαπότηνπιθανότητα $P(F_{(v_1,v_2)} > F_{(v_1,v_2),\alpha}) = \alpha$, δηλ. πάνω από τις οποίες στη δεξιά ουρά της κατανομής υ-πάρχει η πιθανότητα α. Οι βαθμοί ελευθερίας του αριθμητή δίνονται στην πρώτη γραμμή του Πίνακα, ενώ στην πρώτη στήλη αναγράφονται οι βαθμοί ελευθερίας του παρονομαστή.

Επίσης, ακόμα και το MS-Excel έχει δύο συναρτήσεις, αφενός, για τον υπολογισμό των πιθανοτήτων [FDIST(X,df1,df2)] κάτω από την καμπύλη $F_{(v_1,v_2)}$ και αφετέρου, το αντίθετο [FINV(prob.,df_1,df_2)], δηλ. για δεδομένους βαθμούς ελευθερίας οι οποίοι ορίζουν συγκεκριμένη κάθε φορά καμπύλη, και αθροιστική πιθανότητα α δίνει την κρίσιμη τιμή $F_{(v_1,v_2),a}$.

Παράδειγμα 7.10

Να υπολογίσετε τις πιθανότητες $f(F_{(v_1,v_2)})$:

α) $P(F_{(4,10)} > 2,5)$, β) $P(F_{(10,4)} > 2,5)$, γ) $P(F_{(8,21)} > 0,75)$.

Επίσης να βρείτε τις μονοκατάληκτες προς τα πάνω κριτικές τιμές $F_{(v_1,v_2),a}$ για τις οποίες ισχύουν οι παρακάτω πιθανότητες:

δ) $P(F_{(5,15)} > F_{(5,15),0,05}) = 0,05$, ε) $P(F_{(30,6)} > F_{(30,6),0,01}) = 0,01$,

στ) $P(F_{(17,8)} > F_{(17,8),0,025}) = 0,025$.

Απάντηση:

α) $P(F_{(4,10)} > 2,5) = 0,1094$, δηλ. υπάρχει πιθανότητα 10,94% τυχαία να είναι μεγαλύτερη από την κρίσιμη τιμή της $F_{(4,10),0,1094}$=2,5 =2,5

β) $P(F_{(10,4)} > 2,5) = 0,1955$, γ) $P(F_{(8,21)} > 0,75) = 0,6484$.

δ)$F_{(5,15,0,05)} = 2,90$: $P(F_{(5,15)} > 2,90) = 0,05$, δηλ. το μονοκατάληκτο πάνω όριο της $F_{(5,15)}$ έτσι ώστε πάνω από την τιμή αυτή και κάτω από την υπόψη $F_{(5,15)}$ καμπύλη να υπάρχει 5% πιθανότητα (εμβαδόν) είναι η τιμή 2,90.

ε) $F_{(30,6),0,01} = 7,23$: $P(F_{(30,6)} > 7,23) = 0,01$,

στ) $F_{(17,8),0,025} = 4,05$: $P(F_{(17,8)} > 4,05) = 0,025$.

7.3.4 Κατανομή δειγματοληψίας της τυπικής απόκλισης

Κατανομή δειγματοληψίας της τυπικής απόκλισης (s_i) είναι η κατανομή των σχετικών συχνοτήτων των τυπικών αποκλίσεων (s_i, i=1,2,...,k) k ισομεγεθών δειγμάτων μεγέθους (n), τα οποία ελήφθησαν με τυχαία δειγματοληψία, από πληθυσμό μεγέθους N για την τυχαία μεταβλητή X.

Η μορφή αυτής της κατανομής εξαρτάται και εδώ από την κατανομή του πληθυσμού, αν είναι γνωστή ή άγνωστη η πληθυσμιακή διακύμανση (σ^2) και από το μέγεθος του δείγματος (n).

Οι γνωστές θεωρητικές κατανομές πιθανότητας (τυπική κανονική και student-t) που ακολουθεί η στατιστική της τυπικής απόκλισης (s), δίνονται στον παρακάτω συνοπτικό Πίνακα.

Πίνακας 7.5 Κατανομές δειγματοληψίας τυπικής απόκλισης.

σ^2 / Κατανομή Πληθυσμού	Γνωστή	Άγνωστη
Πεπερασμένος Κανονικός	$s \sim N(\sigma, \frac{\sigma^2}{2n}) \Rightarrow$ $z^{*} (\equiv \frac{s-\sigma}{\sigma/\sqrt{2n}}) \sim Z(0,1)$	
Άπειρος Κανονικός		$s \sim t_{\nu}, \quad \nu = (n-1) \Rightarrow$ $t^{*} (\equiv \frac{s-\sigma_0}{s\sqrt{2n}}) \sim t_{\nu}$
Μη-Κανονικός	Άγνωστη $f(s)$	

Όπου σ_0 καθ' υπόθεση γνωστή πληθυσμιακή τυπική απόκλιση.

Παράδειγμα 7.10

Να εκτιμηθεί με πιθανότητα 75% το διάστημα στο οποίο θα βρίσκεται η πληθυσμιακή τυπική απόκλιση (ή μέση απόκλιση τετραγώνου) του χρόνου ολοκλήρωσης ορισμένης παραγωγικής διαδικασίας, η οποία εκ κατασκευής ακολουθεί κανονικό νόμο με μέσο 200 λεπτά και διακύμανση 64 λεπτά, δηλ. X~N(200, 64). Η δειγματοληψία έγινε με δείγμα μεγέθους n=50 ωρών και έδωσε s=6,5 λεπτά.

Απάντηση:

Δεδομένου ότι το μέγεθος του δείγματος είναι μεγάλο η κατανομή δειγματοληψίας της τυπικής απόκλισης ακολουθεί την τυπική κανονική, με μέσο σ και διακύμανση $\sigma^2/2n$ δηλ.:

$s \sim N(\sigma, \frac{\sigma^2}{2n}) \Rightarrow z^*(\equiv \frac{s-\sigma}{\sigma/\sqrt{2n}}) \sim Z(0,1)$. Επομένως το υπόψη διάστημα θα υπολογιστεί υπό την τυπική κανονική καμπύλη με τις γνωστές πιθανότητες, δηλ. από την πιθανότητα:

$$P(-z_{a/2} < Z^* < z_{a/2}) = 0,75 \text{ όπου } Z^* \equiv \frac{s-\sigma}{\sigma/\sqrt{2n}}$$

Αφού λοιπόν $|z_{a/2}| = 1,15$, σ=8, n=50, θα έχουμε:

$$P(6,5-1,15\cdot\frac{8}{\sqrt{2\cdot 50}} < \sigma < 6,5+1,15\cdot\frac{8}{\sqrt{2\cdot 50}}) = 0,75$$

$$5,58 < \sigma < 7,42.$$

Στα 75 από τα 100 δείγματα αναμένουμε η πληθυσμιακή τυπική απόκλιση (σ) να περιέχεται σε διάστημα όπως το παραπάνω.

7.3.5 Κατανομή δειγματοληψίας της αναλογίας

Η δειγματική αναλογία (p) είναι η πιθανότητα «επιτυχίας» σε n διωνυμικές δοκιμές. Το πρόβλημα για τη λύση του οποίου ενδιαφερόμαστε στη διωνυμική κατανομή ήταν η πιθανότητα για τον αριθμό των επιτυχιών X σε n επαναλήψεις του πειράματος Bernoulli, δηλ. για πιθανότητες όπως f(x)=P(X=x).

Αντίθετα, αν ενδιαφερόμαστε μόνο για την αναλογία p και όχι για τις πιθανότητες f(x), τότε εάν από τυχαία δειγματοληψία πάρουμε όλα τα δυνατά ισομεγέθη δείγματα (μεγέθους n) και για κάθε ένα από αυτά υπολογίσουμε και καταγράψουμε σε κατανομή σχετικών συχνοτήτων τις «επιτυχίες» $\hat{p}$ (=x/n (=x/n, αναλογία ή ποσοστό δείγματος), αυτή συνιστά την κατανομή δειγματοληψίας του ποσοστού ή αναλογίας.

Ο μέσος της κατανομής αυτής, από τη διωνυμική γνωρίζουμε ότι είναι Ε($\hat{p}$)=P η άγνωστη πληθυσμιακή αναλογία ή ποσοστό, ενώ η διακύμανσή της είναι $Var(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{P \cdot Q}{n}$, όπου το διωνυμικό Q=1-P. Η τυπική απόκλιση της διακύμανσης αυτής, είναι το τυπικό σφάλμα της κατανομής δειγματοληψίας της αναλογίας, ($\sigma_{\hat{p}} = \sqrt{\frac{P \cdot Q}{n}}$).

Η κατανομή πιθανότητας του $\hat{P}$ για μεγάλα δείγματα είναι η κανονική όπως φαίνεται και στον παρακάτω πίνακα.

Πίνακας 7.6 Κατανομές δειγματοληψίας αναλογίας ή ποσοστού.

Μέγεθος Δείγματος (n) / Καταν.Πληθ.	Μεγάλο $n \geq 30$	Μικρό $n < 30$
Διωνυμικός $X \sim B(n,P)$	$\hat{p} \sim N(P, \frac{PQ}{n}) \Rightarrow$ $Z^{*} (= \frac{\hat{p}-P}{s_{\hat{p}}} = \frac{\hat{p}-P}{\sqrt{\frac{PQ}{n}}}) \sim Z(0,1)$	Διωνυμικός Νόμος και Άβακες (πίνακες διωνυμικής)

Η κατανομή δειγματοληψίας του ποσοστού έχει πολλές εφαρμογές π.χ. στις έρευνες αγοράς προϊόντων-υπηρεσιών, στις δημοσκοπήσεις κοινής γνώμης, στον έλεγχο ποιότητας παραγωγής κ.λπ.

8. Εκπαιδευτική Ενότητα

- **Εκτιμητική και Κλασικοί Παραμετρικοί Έλεγχοι Στατιστικών Υποθέσεων**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατανοεί βασικές έννοιες στους ελέγχους υποθέσεων.
- Να διακρίνει τα διάφορα είδη ελέγχου υποθέσεων.
- Να κατανοεί την έννοια στατιστική σημαντικότητα.
- Να ερμηνεύει τα διαστήματα εμπιστοσύνης.
- Να εφαρμόζει t-test.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Συνέπεια (Consistency)
- Αμεροληψία (Unbiasedness)
- Αποτελεσματικότητα (Efficiency)
- Επάρκεια (Sufficiency)
- Στατιστική Συνάρτηση Ελέγχου
- Περιοχή Απόρριψης-Κρίσιμο Σημείο
- Λάθος Τύπου 1
- Λάθος Τύπου 2
- Ισχύς (1-β)

8.1 Εισαγωγή

Η στατιστική συμπερασματολογία (statistical inference) ολοκληρώνεται σε δύο στάδια. Το πρώτο αφορά στην εκτίμηση της μορφής του πληθυσμού, μέσω των παραμέτρων που τον χαρακτηρίζουν, και το δεύτερο στον έλεγχο στατιστικών υποθέσεων για τις άγνωστες αυτές πληθυσμιακές παραμέτρους.

Τα συμπεράσματα και των δύο σταδίων βασίζονται στην ανάλυση του τυχαίου δείγματος που μελετάμε. Χωρίς τυχαίο δείγμα, με τη στατιστική έννοια του όρου που περιγράψαμε στο προηγούμενο κεφάλαιο, καμία επαγωγή δεν είναι εφικτή. Η γενίκευση (επαγωγή) των συμπερασμάτων του δείγματος, μέσω των τιμών που παίρνουμε από τις στατιστικές του (π.χ. μέσο, διακύμανση, τυπική απόκλιση, αναλογία κ.λπ.), οι οποίες είναι εκτιμήσεις των άγνωστων πληθυσμιακών τιμών των παραμέτρων, δεν είναι δυνατή αν το δείγμα δεν είναι τυχαίο. Μόνο τότε μπορεί να χρησιμοποιηθεί το κύριο σώμα της στατιστικής θεωρίας, δηλ. η πιθανοθεωρία και οι κατανομές δειγματοληψίας, για να επαγάγουμε τα συμπεράσματα του τυχαίου (και ακριβώς γι' αυτό αντιπροσωπευτικού) δείγματος στο γεννήτορα πληθυσμό του.

Η θεωρία της εκτίμησης, στην κλασική επαγωγική στατιστική μεθοδολογία, χρησιμοποιεί τις πληροφορίες του τυχαίου δείγματος με σκοπό να υπολογίσει σε όρους πιθανότητας, δηλ. να εκτιμήσει, την άγνωστη πληθυσμιακή παράμετρο, είτε ως μεμονωμένη τιμή, οπότε μιλάμε για σημειακή εκτίμηση (point estimation), είτε, συνηθέστερα σε διάστημα το οποίο εμπιστευόμαστε με ορισμένη πιθανότητα επαλήθευσης, οπότε μιλάμε για εκτίμηση διαστήματος εμπιστοσύνης (interval estimation).

Η θεωρία του ελέγχου υποθέσεων (hypothesis testing), στην ίδια κλασική μεθοδολογία (όπου δηλ. τα συμπεράσματα για τον άγνωστο πληθυσμό προέρχονται μόνο από τις πληροφορίες του τυχαίου δείγματος), σκοπεύει στον ίδιο με την εκτιμητική στόχο, θέτοντας όμως διαφορετικά το πρόβλημα της εξακρίβωσης της μορφής του πληθυσμού. Ενώ δηλ. στην εκτιμητική θεωρία προσπαθούμε να εντοπίσουμε την «καλύτερη» τιμή ή διάστημα τιμών, μέσα στο οποίο αναμένουμε με ορισμένη πιθανότητα να βρίσκεται η αληθινή αλλά άγνωστη τιμή της παραμέτρου του πληθυσμού που τον χαρακτηρίζει, στον έλεγχο των στατιστικών υποθέσεων εξετάζουμε αν πρέπει να μην αποδεχθούμε μια τιμή ή διάστημα της άγνωστης πληθυσμιακής παραμέτρου. Αυτό μάλιστα που είναι ιδιαίτερα σημαντικό να έχουμε κατανοήσει στη θεωρία του κλασικού παραμετρικού ελέγχου στατιστικών υποθέσεων είναι ότι τελικώς ελέγχουμε αν «το τυχαίο δείγμα μας προέρχεται από τον πληθυσμό από τον οποίο πιστεύουμε ότι το αντλήσαμε»;

Τα επόμενα δύο τμήματα είναι αφιερωμένα στην εκτιμητική, παραθέτοντας όπως μέχρι τώρα, στοιχεία στατιστικής θεωρίας και δίνοντας έμφαση στις εφαρμογές, ενώ το τελευταίο, στην ίδια λογική αφιερώνεται στον κλασικό παραμετρικό έλεγχο στατιστικών υποθέσεων.

8.2 Εκτιμητές και μέθοδοι εκτίμησης

8.2.1 Εκτιμητές και ιδιότητές τους

Ήδη από την μέχρι τώρα παρουσίαση πρέπει να έχει γίνει σαφής, στον προσεκτικό αναγνώστη, η σημαντική διάκριση μεταξύ εκτιμητή ή εκτιμήτριας (estimator) ή στατιστικής συνάρτησης, από τη μια μεριά, και εκτίμησης (estimation), από την άλλη. Εκτιμητής του πληθυσμιακού μέσου μ είναι, για παράδειγμα, ο δειγματικός $\bar{X}$ ή αντίστοιχα εκτιμήτρια

της πληθυσμιακής διακύμανσης σ^2 μπορεί να είναι η δειγματική s^2. Σημειώστε επίσης τη διάκριση στη γραφή, ανάμεσα στις παραμέτρους του πληθυσμού που συμβολίζονται με ελληνικά γράμματα, και τις στατιστικές του δείγματος, οι οποίες γράφονται με λατινικά γράμματα.

Γνωρίζουμε πλέον ότι το χαρακτηριστικό του στατιστικού πληθυσμού, το οποίο μας ενδιαφέρει, το συμβολίζουμε με την τυχαία μεταβλητή, έστω, Χ. Η πιθανοθεωρητική συμπεριφορά της Χ περιγράφεται από τη συνάρτηση πιθανότητάς της, έστω, f(x) ή αναλυτικότερα $f(x;\Theta_j),\ j=1,2,\ldots,k$, συμβολισμός που τονίζει ότι η υπόψη πιθανοθεωρητική κατανομή της Χ προσδιορίζεται πλήρως από τις j παραμέτρους Θ(j). Έτσι, για παράδειγμα για μια κανονική ή Gauss-Laplace Χ θα γράφαμε $f(x;\Theta_1,\Theta_2)$, όπου $(\Theta_1 \equiv \mu)$ και $(\Theta_2 \equiv \sigma^2)$, ή $f(x;\mu,\sigma^2)$,, αφού η κανονική περιγράφεται πλήρως από το μέσο και τη διακύμανσή της.

Το πρόβλημα λοιπόν της εκτίμησης, έστω π.χ. για την τυχαία κανονική Χ μεταβλητή με συνάρτηση πιθανότητας $f(x;\mu,\sigma^2)$, μπορεί να τεθεί ως εξής: «από στατιστικές συναρτήσεις των παρατηρήσεων τυχαίου δείγματος μεγέθους (n), το οποίο αντλήσατε από τον πληθυσμό $f(x;\mu,\sigma^2)$, να υπολογίσετε σε όρους πιθανότητας,δηλ. να εκτιμήσετε, τις άγνωστες πληθυσμιακές παραμέτρους του μέσου (μ) και της διακύμανσης (σ^2), οι οποίες τον περιγράφουν πλήρως». Με βάση την ακρίβεια (accuracy) της γνώσης για τις τιμές των πληθυσμιακών παραμέτρων (εκτιμήσεις), η οποία δίνεται σε όρους πιθανότητας, θα είναι δυνατόν, να μιλήσουμε για τη συγκεκριμένη μορφή της πληθυσμιακής κατανομής (καμπύλη πιθανότητας), και επομένως, να κάνουμε προβλέψεις για τη γενική συμπεριφορά του φαινομένου Χ.

Κατά συνέπεια, κατ' αρχάς, το πρόβλημα της εκτίμησης εστιάζεται στην επιλογή της «κατάλληλης» στατιστικής ή εκτιμήτριας, η οποία επαναλαμβάνουμε είναι κάποια συνάρτηση των τυχαίων μεταβλητών-παρατηρήσεων του δείγματος. Έτσι, ο εκτιμητής $(\hat{\theta}_j)$ της αντίστοιχης άγνωστης πληθυσμιακής παραμέτρου Θ(j) παριστάνεται γενικά ως εξής:

$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$	**(8.1)**

Όπου n το μέγεθος του τυχαίου δείγματος για τη Χ με τιμές (x_i), i=1,2,3,...n.

Το πόσο καλός είναι ο εκτιμητής $(\hat{\theta}_j)$ για να μας δώσει «ακριβή» εκτίμηση της άγνωστης πληθυσμιακής παραμέτρου, εφαρμόζοντας πάνω του της τυχαίες παρατηρήσεις του δείγματος, και μάλιστα εκτίμηση, είτε σημειακή δηλ. συγκεκριμένη τιμή, είτε, σε διάστημα, το οποίο πριν την εκτίμηση, εμπιστευόμαστε με ορισμένη πιθανότητα, εξαρτάται από τις ιδιότητες που τον χαρακτηρίζουν. Με άλλα λόγια, τα κριτήρια επιλογής εκτιμητών, έτσι ώστε να πετύχουμε τη μέγιστη ποιότητα των εκτιμήσεων αναφέρονται στις επιθυμητές ιδιότητες που αυτοί πρέπει να έχουν.

Οι ιδιότητες αυτές διακρίνονται για μικρά και για μεγάλα δείγματα, ενώ εκείνοι οι εκτιμητές που τις έχουν ονομάζονται άριστοι (best estimators) αφού οδηγούν σε άριστες εκτιμήσεις. Με απλή αναφορά τους[9] παρακάτω, οι ιδιότητες, σε μικρά δείγματα, που πρέπει να έχει κάποιος εκτιμητής για να χαρακτηριστεί άριστος είναι: α) η αμεροληψία, β) η αποτελεσματικότητα και γ) η επάρκεια.

• Ο εκτιμητής $\hat{\theta}$ είναι αμερόληπτος της άγνωστης πληθυσμιακής παραμέτρου Θ εάν η προσδοκώμενη τιμή του είναι ίση με την παράμετρο. Συμβολικά γράφουμε για την ιδιότητα της αμεροληψίας (unbiasedness) του $\hat{\theta}$, γενικά ως εξής:

$E(\hat{\theta}) = \Theta$	**(8.2)**

$\text{μερ.σφ} = E(\hat{\theta}) - \Theta = E(\hat{\theta} - \Theta)$ ενώ $\text{δ.σφ.} = \hat{\theta} - \Theta$	**(8.3)**

Όπου:
--- «μερ.σφ.» είναι το μεροληπτικό σφάλμα (bias error) που ορίζεται ως η απόκλιση της αναμενόμενης τιμής της εκτιμήτριας από την αντίστοιχη παράμετρο της, και

--- «δ.σφ.» είναι το δειγματοληπτικό σφάλμα (sampling error) που ορίζεται ως η απόκλιση της εκτίμησης από την αντίστοιχη τιμή της άγνωστης παραμέτρου της.

Για παράδειγμα, αποδεικνύεται ότι, η αναμενόμενη τιμή της κατανομής δειγματοληψίας του μέσου είναι ο πληθυσμιακός μέσος ή $E(\bar{X}) = \mu$, όπου $\bar{X} = \frac{1}{n}\sum_i x_i$. Επομένως ο δειγματικός μέσος είναι αμερόληπτος εκτιμητής του μέσου μ. Επίσης, η δειγματική διακύμανση $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{X})^2$ είναι αμερόληπτη εκτιμήτρια της πληθυσμιακής σ^2, ενώ αντίθετα δεν συμβαίνει το ίδιο και με την τυπική απόκλιση.

• Εντούτοις, η αμεροληψία δεν αρκεί από μόνη της για την επιλογή του άριστου εκτιμητή. Ένας αμερόληπτος εκτιμητής με μεγάλη διακύμανση, και αντίστροφα, ένας εκτιμητής με ελάχιστη διακύμανση αλλά μεροληπτικός, δεν είναι επιθυμητοί. Ο συνδυασμός αμεροληψίας και ελάχιστης διακύμανσης μας οδηγεί στη δεύτερη σημαντική ιδιότητα των εκτιμητών, την αποτελεσματικότητα (efficiency).

[9] Υπενθυμίζεται ότι δεν είναι στους στόχους του παρόντος η απόδειξη των σχέσεων, η οποία όμως συνιστάται στους αναγνώστες να διερευνηθεί στα σχετικά στατιστικά εγχειρίδια, ως η καλύτερη άσκηση για την εμπέδωση της θεωρίας, χωρίς την οποία οι εφαρμογές φαντάζουν «μηχανιστικές». Ενδεικτικά από αυτά τα εγχειρίδια αναφέρονται στη βιβλιογραφία.

Γενικά, ορίζουμε ως αποτελεσματικό τον εκτιμητή $\hat{\theta}$ της αντίστοιχης πληθυσμιακής παραμέτρου Θ, εάν μεταξύ των αμερόληπτων έχει τη μικρότερη διακύμανση, δηλ. εάν ικανοποιεί τις σχέσεις:

1) $\hat{\theta}$ αμερόληπτος και 2) $Var(\hat{\theta}) < Var(\theta')$, όπου θ' οποιοσδήποτε άλλος αμερόληπτος	**(8.4)**

Η επιλογή του πλέον αποτελεσματικού εκτιμητή ή άριστου αμερόληπτου εκτιμητή (best unbiased estimator) μέσα από το σύνολο των αμερόληπτων εκτιμητών είναι προφανώς πολύ δύσκολη εργασία. Ο όγκος της τελευταίας περιορίζεται αν επιλέξουμε από τους αμερόληπτους μόνο εκείνους που είναι γραμμικές συναρτήσεις των παρατηρήσεων του δείγματος. Έτσι φτάνουμε στην έννοια του άριστου γραμμικού αμερόληπτου εκτιμητή (best linear unbiased estimator, ή BLUE) που είναι εκείνος που ικανοποιεί τις προϋποθέσεις,

1) $\hat{\theta}$ γραμμική συνάρτηση των x_i 2) $\hat{\theta}$ αμερόληπτος και 3) $Var(\hat{\theta}) < Var(\theta')$, όπου θ' οποιοσδήποτε άλλος αμερόληπτος	**(8.5)**

Σημειώνεται ότι πολλές συναρτήσεις των δειγματικών τιμών xi μπορούν να προσεγγιστούν γραμμικά και επομένως η ιδιότητα BLUE δεν είναι περιοριστική.

- Ο εκτιμητής $\hat{\theta}$ είναι επαρκής της άγνωστης πληθυσμιακής παραμέτρου Θ, εάν χρησιμοποιεί όλες τις διαθέσιμες παρατηρήσεις του τυχαίου δείγματος. Η επάρκεια (sufciency), αναγκαία συνθήκη για την αποτελεσματικότητα, είναι ιδιότητα που διαθέτει για παράδειγμα, ο δειγματικός μέσος $\bar{X} = \frac{1}{n}\sum_i x_i$, αλλά όχι και η διάμεσος $Me = \frac{n-1}{2}_{\theta\acute{\varepsilon}\sigma\eta}$ των κατ' αύξουσα διάταξη ταξινομημένων παρατηρήσεων x_i.

Επιπλέον, οι ιδιότητες των στατιστικών ή εκτιμητριών, για τις οποίες κατά την προσπάθεια κατασκευής των κατανομών δειγματοληψίας τους, χρησιμοποιούμε όλο και μεγαλύτερα μεγέθη δειγμάτων, ή με αυστηρή διατύπωση, το μέγεθος n του τυχαίου δείγματος τείνει στο άπειρο, ονομάζονται ασυμπτωτικές ιδιότητες. Οι τελευταίες πήραν το όνομά τους από τις ασυμπτωτικές κατανομές δειγματοληψίας των εκτιμητριών, οι οποίες προκύπτουν από τις οριακές (marginal) κατανομές τους, όταν δηλ. το μέγεθος του δείγματος τείνει στο άπειρο. Οι ιδιότητες αυτές είναι α) η ασυμπτωτική αμεροληψία, β) η συνέπεια και γ) η ασυμπτωτική αποτελεσματικότητα.

- Εάν η αναμενόμενη τιμή του εκτιμητή $\hat{\theta}$ «βρίσκει» την άγνωστη πληθυσμιακή παράμετρο Θ, καθώς το μέγεθος του χρησιμοποιούμενου τυχαίου δείγματος συνεχώς μεγαλώνει, τότε λέμε ότι έχει την ιδιότητα της ασυμπτωτικής αμεροληψίας (asymptotic unbiasness). Η τελευταία συμβολικά δίνεται από τη σχέση:

$\lim_{n\to\infty} E(\hat{\theta}) = \Theta$	**(8.6)**

- Συνεπής (consistent) ονομάζεται ο εκτιμητής $\hat{\theta}$, ο οποίος συγκεντρώνεται ολοένα και περισσότερο γύρω από την αντίστοιχη άγνωστη πληθυσμιακή παράμετρο Θ όσο το μέγεθος n του δείγματος τείνει στο άπειρο. Συμβολικά η έννοια της συνέπειας δίνεται συνήθως με τη χρήση του μέσου σφάλματος τετραγώνου (mean square error, MSE που μετράει τη διασπορά των εκτιμήσεων γύρω από την αληθινή παράμετρο του πληθυσμού), ως εξής:

$p\lim_{n\to\infty} MSE(\hat{\theta}) = 0$ ή $p\lim_{n\to\infty} (\hat{\theta}) = \Theta$	**(8.7)**

$MSE(\hat{\theta}) = E(\hat{\theta}-\Theta)^2$ $= Var(\hat{\theta}) + bias(\hat{\theta})^2$	**(8.8)**

Όπου $p\lim_{n\to\infty} \hat{\theta} = \theta^*$ είναι το όριο πιθανότητας του εκτιμητή $\hat{\theta}$, δηλ. το σημείο θ^* στο οποίο συγκεντρώνεται η κατανομή δειγματοληψίας του $\hat{\theta}$ όταν το μέγεθος των τυχαίων δειγμάτων τείνει στο άπειρο.

- Τέλος, ασυμπτωτικά αποτελεσματικός (asymptotic efficient) είναι ο εκτιμητής $\hat{\theta}$ όταν από όλους τους συνεπείς έχει τη μικρότερη ασυμπτωτική διακύμανση.

8.2.2 Μέθοδοι εύρεσης εκτιμητών πληθυσμιακών παραμέτρων

Οι θεωρητικοί της στατιστικής επιστήμης έχουν αναπτύξει μεθόδους εύρεσης εκτιμητών, οι οποίες χρησιμοποιώντας τεχνικές βελτιστοποίησης πετυχαίνουν να ικανοποιούν τις περισσότερες (αν όχι όλες) από τις παραπάνω αναφερθείσες επιθυμητές τους ιδιότητες.

Έτσι δεν χρειάζεται πλέον, για συνήθη πρακτικά προβλήματα, να ψάχνουμε τον άριστο αμερόληπτο ή BLUE εκτιμητή που είναι κατάλληλος για την ειδική περίπτωση της συμπεριφοράς που εξετάζουμε.

Οι κλασικές μέθοδοι εύρεσης εκτιμητών, των αγνώστων αλλά αληθινών παραμέτρων του δειγματοληπτούμενου πληθυσμού, είναι βασικά πέντε:

1. η μέθοδος των ελαχίστων τετραγώνων (least squares' method)
2. η μέθοδος της μέγιστης πιθανοφάνειας (maximum likelihood method)
3. η μέθοδος των ροπών (method of moments)
4. η μέθοδος Bayes (Bayesian approach) και
5. η μέθοδος της άριστης γραμμικής αμερόληπτης εκτίμησης (best linear unbi-asedness method).

Ακόμα και η απλούστερη δυνατή παρουσίαση της μεθοδολογίας των τεχνικών αυτών απαιτεί διαφορετικό τρόπο προσέγγισης από αυτόν της εφαρμοσμένης στατιστικής που έχουμε υιοθετήσει στο βιβλίο αυτό. Ο αναγνώστης που επιθυμεί να εμβαθύνει στην πολύ ενδιαφέρουσα αυτή στατιστική θεωρία, προτρέπεται να μελετήσει σχετικά εγχειρίδια, από τα οποία ορισμένα ενδεικτικά αναφέρονται στη βιβλιογραφία.

8.3 Εκτίμηση σε διάστημα εμπιστοσύνης

8.3.1 Έννοια και μεθοδολογία κατασκευής διαστημάτων εμπιστοσύνης

Εάν με βάση τις παραπάνω ιδιότητες των εκτιμητών, επιλέξουμε έναν $\hat{\theta}$ π.χ. BLUE, και εφαρμόσουμε σ' αυτόν τα στοιχεία xi τυχαίου δείγματος μεγέθους n, τότε η τιμή που παίρνουμε αποτελεί τη σημειακή εκτίμηση για την άγνωστη αλλά αληθινή πληθυσμιακή. Βέβαια, δεν περιμένουμε ποτέ η σημειακή εκτίμηση να «πετύχει» ακριβώς την πληθυσμιακή. Εντούτοις, όπως συμπεραίνεται εύκολα από τις επιθυμητές ιδιότητες των εκτιμητών που έχουν αποδείξει οι στατιστικολόγοι και αναφέραμε παραπάνω, η ακρίβεια της εκτίμησης μας θα αυξάνει συνεχώς, όσο το μέγεθος του τυχαίου δείγματος επίσης αυξάνει, δεδομένης της επιλογής «άριστου» εκτιμητή.

Επιπλέον, στην προσπάθεια μας να εκτιμήσουμε τις άγνωστες παραμέτρους που περιγράφουν τη δομή του πληθυσμού, με βάση την πληροφόρηση του δείγματος, είναι προτιμότερο (αφού έτσι αυξάνουμε την αξιοπιστία της εκτίμησης) να επιλέξουμε εκτίμηση σε διάστημα εμπιστοσύνης με προκαθορισμένη πιθανότητα (1-a) σωστής επιλογής. Όπου a εκφράζει μια μορφή λάθους απόφασης.

Εδώ λοιπόν, το πρόβλημα μας εντοπίζεται στην εύρεση των ορίων (l, u) του διαστήματος έστω δ. Πιο συγκεκριμένα τα δύο όρια που ψάχνουμε πρέπει να δίνονται από την πιθανότητα:

$P(l \leq \Theta \leq u) = 1 - \alpha \Leftrightarrow$ $P[\Theta \notin (l, u)] = \alpha$	**(8.9)**

Όπου 0≤a≤1 εκφράζει μια μορφή λανθασμένης απόφασης (decision).

Επομένως, η (8.9) μας δίνει την εκτίμηση σε διάστημα εμπιστοσύνης της Θ, δηλ. τα όρια [l, u], με πιθανότητα 1-α. Η πιθανότητα αυτή 1-α ονομάζεται επίπεδο εμπιστοσύνης (confidence level), ενώ το συμπλήρωμά της α είναι γνωστή ως επίπεδο σημαντικότητας (level of significance).

Για παράδειγμα, εάν από τα στοιχεία τυχαίου δείγματος υπολογίσαμε το μέσο $\bar{X}$, επειδή αποδεικνύεται ότι αυτός είναι BLUE εκτιμητής του άγνωστου πληθυσμιακού μ, τότε η τιμή του δειγματικού μας μέσου, έστω, $\bar{X}_i$ αποτελεί σημειακή εκτίμηση για την αντίστοιχη παράμετρο του πληθυσμού μ.

Η εκτίμηση σε διάστημα εμπιστοσύνης για την ίδια παράμετρο μ χρησιμοποιεί, αφενός, τη σημειακή εκτίμηση (επομένως στηρίζεται πάνω στις ιδιότητες των εκτιμητών), και αφετέρου, την κατανομή δειγματοληψίας του μέσου.

Πρέπει να τονιστεί, ότι επειδή όλες οι εκτιμήτριες είναι συναρτήσεις τυχαίων μεταβλητών (x_i), είναι προφανές ότι και αυτές είναι τυχαίες μεταβλητές. Κατά συνέπεια, τόσο η σημειακή εκτίμηση $\hat{\theta}_i$ όσο και τα όρια εμπιστοσύνης [l, u], είναι επίσης τυχαίες μεταβλητές, οι οποίες προσδιορίζονται από την κατανομή δειγματοληψίας της $\hat{\theta}$. Η γνώση της κατανομής δειγματοληψίας όμως, μας δίνει τη δυνατότητα να μπορούμε να προσδιορίσουμε το βαθμό εμπιστοσύνης 1-α των όποιων εκτιμήσεων μας.

Για να διευκολύνουμε την παρουσίαση της μεθοδολογίας εύρεσης του διαστήματος εμπιστοσύνης άγνωστης πληθυσμιακής παραμέτρου, θα υποθέσουμε τη συνηθέστερη περίπτωση, ότι δηλ. η παράμετρος που μας ενδιαφέρει να εκτιμήσουμε σε διάστημα είναι ο μέσος μ. Έστω επίσης, ότι το μέγεθος του τυχαίου μας δείγματος από τον άπειρο και κανονικό πληθυσμό που δειγματοληπτούμε είναι μεγάλο (n>30), ενώ, είναι γνωστή και η πληθυσμιακή διακύμανση σ^2. Οι υποθέσεις αυτές περιγράφουν το τυχαίο μας πείραμα με ακρίβεια, έτσι ώστε να μας οδηγήσουν στην κατάλληλη κατανομή δειγματοληψίας που πρέπει να χρησιμοποιήσουμε.

Το κυριότερο «εργαλείο» ανάλυσης που μας προσφέρει η στατιστική θεωρία για να λύσουμε το πρόβλημα της εκτίμησης διαστήματος εμπιστοσύνης, είναι η γνώση της κατανομής δειγματοληψίας του μέσου $\bar{X}$, τον οποίο χρησιμοποιούμε ως BLUE εκτιμητή του άγνωστου πληθυσμιακού μ.

Στους Πίνακες 7.3 και 7.4 περιγράψαμε τις δυνατές περιπτώσεις των κατανομών δειγματοληψίας του μέσου $\bar{X}$ ανάλογα με τη μορφή του γεννήτορα πληθυσμού, τη γνωστή ή όχι πληθυσμιακή διακύμανση σ^2, και το μέγεθος του τυχαίου μας δείγματος. Επομένως με βάση τα δεδομένα του παρόντος τυχαίου πειράματος και τη γνώση από τους υπόψη Πίνακες προκύπτει ότι εδώ έχουμε την περίπτωση ο δειγματικός μέσος να ακολουθεί κανονική με μέσο τον πληθυσμιακό μ και διακύμανση σ^2/n,

δηλ. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Κατά συνέπεια, η τυποποιημένη μεταβλητή του $\bar{X}$, κάποια Z* θα είναι προφανώς με απλή αντικατάσταση στον τυπικό μετασχηματισμό $\frac{X-\mu}{\sigma}$ η $Z^* \equiv \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \sim Z(0,1)$, δηλ. η σχέση (7.6).

Η τελευταία αυτή συνάρτηση βλέπουμε ότι περιλαμβάνει τόσο πληροφόρηση από το δείγμα μας ($\bar{X}$, n) ή διαθέσιμη (σ^2), όσο και τη ζητούμενη άγνωστη πληθυσμιακή παράμετρο μ.

Αυτήν τη συνάρτηση (7.6) αν αντικαταστήσουμε στο διάστημα εμπιστοσύνης (8.9) που αποτελεί το γενικό κανόνα, θα βρούμε το ζητούμενο διάστημα εμπιστοσύνης για το μέσο μ. Αναλυτικότερα θα έχουμε:

- $P(l \le \Theta \le u) = 1-\alpha \Rightarrow P(-z_{\alpha/2} \le Z^* \le z_{\alpha/2}) = 1-\alpha \Rightarrow P(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \le z_{\alpha/2}) = 1-\alpha$

αλλά $(\sigma/\sqrt{n}) \equiv \sigma_{\bar{X}}$ δηλ. το τυπικό σφάλμα του μέσου, ή τυπική απόκλιση της κατανομής δειγματοληψίας του $\bar{X}$. Επομένως με αντικατάσταση για απλοποίηση θα έχουμε:

- $P(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \le z_{\alpha/2}) = 1-\alpha \Rightarrow P(\bar{X} - z_{\alpha/2}\cdot\sigma_{\bar{X}} \le \mu \le \bar{X} + z_{\alpha/2}\cdot\sigma_{\bar{X}}) = 1-\alpha$

Άρα το ζητούμενο διάστημα εμπιστοσύνης του μέσου μ με πιθανότητα εμπιστοσύνης 1-α θα δίνεται από την ανισότητα:

$\bar{X} - z_{\alpha/2}\cdot\sigma_{\bar{X}} \le \mu \le \bar{X} + z_{\alpha/2}\cdot\sigma_{\bar{X}}$	**(8.10)**

Με άλλα λόγια, τα όρια εμπιστοσύνης του μέσου μ που μας έδωσε η εκτίμηση του διαστήματος εμπιστοσύνης (8.10) με πιθανότητα 1-α είναι:

$(l,u) = \{(\bar{X} - z_{\alpha/2}\cdot\sigma_{\bar{X}}), (\bar{X} + z_{\alpha/2}\cdot\sigma_{\bar{X}})\}$ ή $(l,u) = (\bar{X} \pm z_{\alpha/2}\cdot\sigma_{\bar{X}})$	**(8.11)**

Συμπεραίνουμε λοιπόν ότι, γενικός κανόνας για την εκτίμηση των διαστημάτων εμπιστοσύνης είναι ο εξής:

Σημειακή Εκτίμηση ± (κρίσιμη τιμή x τυπικό σφάλμα εκτίμησης)	**(8.12)**

Από το διάστημα (8.11) βλέπουμε ότι η ακρίβεια της εκτίμησης εκφράζεται από το πλάτος του διαστήματος δ το οποίο με τη σειρά του προσδιορίζεται από την ποσότητα $\delta = z_{\alpha/2} \cdot \sigma_{\bar{X}} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$. Άρα η ακρίβεια της εκτίμησης εξαρτάται, τόσο, από το επίπεδο εμπιστοσύνης 1-α που προσδιορίζει τις κρίσιμες τιμές $z_{\alpha/2}$, όσο και, από το μέγεθος του τυπικού σφάλματος, το οποίο συσχετίζεται αρνητικά με το μέγεθος του δείγματος n. Οι αποτελεσματικοί εκτιμητές φαίνεται εδώ ότι είναι ιδιαίτερα χρήσιμοι.

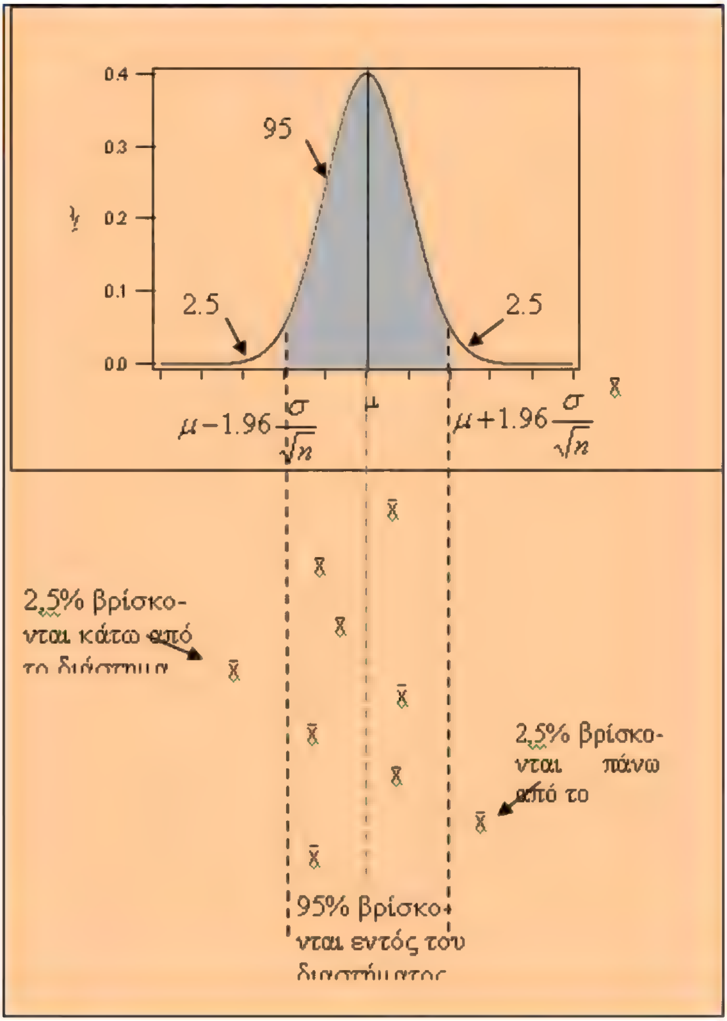
Συνοψίζοντας, η διαδικασία εύρεσης ορίων εμπιστοσύνης για την εκτίμηση σε διάστημα πληθυσμιακών παραμέτρων ακολουθεί τα βήματα του παρακάτω πίνακα.

Πίνακας 8.1 Μεθοδολογία εκτίμησης παραμέτρων σε διαστήματα εμπιστοσύνης

Βήμα 1ο	**Εκτιμήτρια**. Επιλέγουμε με βάση τις επιθυμητές ιδιότητες των εκτιμητών, την εκτιμήτρια της άγνωστης πληθυσμιακής παραμέτρου Θ, έστω για παράδειγμα $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ στην περίπτωση του μέσου μ. Υπολογίζουμε την τιμή της $\bar{X}$ (σημειακή εκτίμηση) και το τυπικό της σφάλμα ($\sigma_{\bar{X}}$ ή $s_{\bar{X}}$).
Βήμα 2ο	**Στατιστική ελέγχου**, έστω, $Y = f(\hat{\theta}, \Theta)$. Επιλέγουμε τυχαία μεταβλητή (αν είναι δυνατόν τυποποιημένη), η οποία να είναι συνάρτηση τόσο της εκτιμήτριας όσο και της παραμέτρου, δηλ. $Y = f(\bar{X}, \mu)$. Στην περίπτωση του μέσου, υπό τις γνωστές προϋποθέσεις (βλ. Πίνακες 7.3 & 7.4) αυτή μπορεί να είναι η τυποποιημένη $Z^{*} (\equiv \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}) \sim Z(0,1)$.
Βήμα 3ο	**Επίπεδο εμπιστοσύνης, κρίσιμες τιμές**. Επιλέγουμε επίπεδο εμπιστοσνης 1-α που προσδιορίζει κριτικές τιμές, π.χ. εδώ για το μέσο οι κρίσιμες τιμές για επίπεδο σημαντικότητας *α* θα είναι $\lvert z_{\alpha/2} \rvert$.
Βήμα 4ο	**Όρια εμπιστοσύνης**. Υπολογίζουμε τα όρια εμπιστοσύνης για το επίπεδο 1-α, βάσει της πιθανότητας $P(l<\Theta<u)=1-\alpha$, αντικαθιστώντας τη Θ με τη στατιστική της σχετικής κατανομής δειγματοληψίας (βήμα 2ο). Εναλλακτιά εφαρμόζουμε το γενικό κανόνα (8.12).

Στο επόμενο Σχήμα φαίνεται παραστατικά η εκτίμηση διαστήματος εμπιστοσύνης του μέσου.

Σχήμα 8.1 Κατανομή δειγματοληψίας μέσου



8.3.2 Εφαρμογές: Διαστήματα εμπιστοσύνης για το μέσο, την τυπική απόκλιση και την αναλογία

Κατ' αρχάς να επισημάνουμε ότι οι εκτιμήσεις διαστημάτων εμπιστοσύνης που παρουσιάζουμε εδώ, αναφέρονται σε μονομεταβλητούς πληθυσμούς και επομένως στη χρήση ενός (1) τυχαίου δείγματος.

Η περίπτωση της εκτίμησης του πληθυσμιακού μέσου αριθμητικού μ σε διάστημα εμπιστοσύνης 1-α είναι η συνηθέστερα χρησιμοποιούμενη στην πράξη. Η εκτίμηση του μέσου προσδιορίζεται από την κατανομή δειγματοληψίας του $\bar{X}$, την οποία συνοπτικά παρουσιάσαμε στους Πίνακες 7.3 και 7.4. Αυτοί μας δίνουν την απαραίτητη γνώση της πιθανοθεωρητικής συμπεριφοράς της εκτιμήτριας μας $\bar{X}$. Εφαρμόζοντας λοιπόν τη γνώση αυτή στη διαδικασία εκτίμησης παραμέτρων σε διάστημα εμπιστοσύνης την οποία, επίσης συνοπτικά, περιγράφουμε στον Πίνακα 8.1 έχουμε τη λύση του προβλήματος.

Το μόνο στοιχείο που αξίζει να επισημανθεί επιπλέον εδώ, είναι ότι η περίπτωση της κατανομής δειγματοληψίας του μέσου διακρίνεται, όπως είναι γνωστό, ανάλογα με **α)** τη μορφή του πληθυσμού, **β)** το αν είναι ή όχι γνωστή η πληθυσμιακή διακύμανση σ^2, και **γ)** το μέγεθος n του τυχαίου που χρησιμοποιούμε. Στις περισσότερες περιπτώσεις όμως χρησιμοποιούμε την τεράστιας σημασίας ανακάλυψη των θεωρητικών της στατιστικής, το **Κεντρικό Οριακό Θεώρημα**, στο οποίο τονίζεται ότι ανεξάρτητα από τη μορφή του γεννήτορα πληθυσμού του δείγματός μας, όσο το μέγεθός του n αυξάνει, η κατανομή δειγματοληψίας του μέσου $\bar{X}$ προσεγγίζει την κανονική και του τυποποιημένου δειγματικού μέσου, την τυπική κανονική.

Με άλλα λόγια, όταν χρησιμοποιούμε μεγάλα τυχαία δείγματα (n≥30), οι τυχαίες μεταβλητές που θα χρησιμοποιούμε, τόσο, στις εκτιμήσεις διαστημάτων εμπιστοσύνης, όσο, και στους αντίστοιχους ελέγχους υποθέσεων (βλ. το επόμενο τμήμα), θα είναι η $Z^*(\equiv \frac{\bar{X}-\mu}{(s/\sqrt{n})}) \sim Z(0,1)$. Στην περίπτωση δε των μικρών τυχαίων δειγμάτων (n<30), όπου εργαζόμαστε κάτω από το νόμο του Gosset ή t-Student, θα χρησιμοποιούμε την $t^*(\equiv \frac{\bar{X}-\mu}{(s/\sqrt{n})}) \sim t_\nu, \quad \nu = n - .$

Τα παραπάνω θα γίνουν σαφέστερα με τα παραδείγματα που ακολουθούν.

Παράδειγμα 8.1

Βιομηχανία συσκευασίας γάλακτος θέλει να ελέγξει την αξιοπιστία νέων μηχανημάτων που επιθυμεί να αγοράσει. Προκειμένου να ελέγξει τις προδιαγραφές τους, ο διευθυντής παραγωγής πήρε τυχαίο δείγμα 25 κουτιών και βρήκε μέσο 492 και διακύμανση 121. Ποια είναι η εκτίμησή σας για τον πληθυσμιακό μέσο των μηχανημάτων, αν αφενός, γνωρίζετε ότι το δείγμα ήταν πράγματι απλό τυχαίο, και αφετέρου, δέχεστε επίπεδο σημαντικότητας α=5%.

Απάντηση:

Εξαιτίας του μικρού δείγματος n=25 μπορεί να χρησιμοποιηθεί η t-Student κατανομή δειγματοληψίας. Επομένως εφαρμόζοντας τα 4 στάδια της διαδικασίας του Πίνακα 8.1 για την εκτίμηση διαστήματος εμπιστοσύνης θα έχουμε:

Βήμα 1ο

Εκτιμήτρια. Η εκτιμήτρια της άγνωστης πληθυσμιακής παραμέτρου μ είναι $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ την οποία χρησιμοποίησε ο εν λόγω διευθυντής παραγωγής και πήρε ως σημειακή εκτίμηση $\bar{X}$=492 . Το τυπικό της σφάλμα είναι $s_{\bar{X}} = \sqrt{\frac{121}{25}} = 2,20$.

Βήμα 2ο

Στατιστική ελέγχου, $Y = f(\hat{\theta}, \Theta)$. Η τυχαία μεταβλητή και μάλιστα του τυποποιημένου μέσου, που μας υποδεικνύει η κατανομή δειγματοληψίας στη συγκεκριμένη περίπτωση (άγνωστη πληθυσμιακή διακύμανση σ^2, και μικρό δείγμα) είναι συνάρτηση τόσο της εκτιμήτριας όσο και της παραμέτρου, δηλ. $t^*(\equiv \frac{\bar{X}-\mu}{s_{\bar{X}}}) \sim t_\nu, \quad \nu = n-1$..

Βήμα 3ο

Επίπεδο εμπιστοσύνης, κρίσιμες τιμές. Το επιλεγμένο επίπεδο εμπιστοσύνης 1-α είναι 1-0,05=0,95 ή 95% και προσδιορίζει κριτικές τιμές, τις οποίες βρίσκουμε από τους Πίνακες της t-Student, στην $t_{25-1} = t_{24}$, τις $|t_{24,0,05}| = 2,064$..

Βήμα 4ο

Όρια εμπιστοσύνης. Υπολογίζουμε τα όρια εμπιστοσύνης 95%, βάση της πιθανότητας P(l<Θ<u)=95% $P(|t^*| < t_{\nu,\alpha/2}) = 95\%$ ή $P(-t_{\nu,\alpha/2} < t^* < t_{\nu,\alpha/2}) = 95\%$

ή $P(\bar{X} - t_{\nu,\alpha/2} \cdot s_{\bar{X}} < \mu < \bar{X} + t_{\nu,\alpha/2} \cdot s_{\bar{X}}) = 95\%$. Επομένως το ζητούμενο 95% διάστημα εμπιστοσύνης θα είναι $\bar{X} \pm t_{\nu,\alpha/2} \cdot s_{\bar{X}} = 492 \pm 2,064 \cdot 2,20 = (487,46, 496,54)$ ή 487,46 < μ < 496,54.

Η ερμηνεία της εκτίμησης αυτής απαιτεί ιδιαίτερη προσοχή. Γενικά μπορούμε να πούμε ότι, σε επαναλαμβανόμενες τυχαίες δειγματοληψίες από τον υπόψη άπειρο πληθυσμό το 95% των δειγμάτων θα δώσουν διάστημα στο οποίο θα περιέχεται η αληθινή αλλά άγνωστη παράμετρος μ. Ένα από αυτά τα δείγματα είναι και το δικό μας από το οποίο πήραμε εκτίμηση στο διάστημα (487,46, 496,54) για το μέσο βάρος. Ο μέσος του πληθυσμού μ είτε θα είναι εντός του διαστήματος (487,46, 496,54) είτε δεν θα είναι.

Επομένως είναι λάθος η έκφραση «με βαθμό εμπιστοσύνης 95% το διάστημα το οποίο θα περιλαμβάνει τον άγνωστο μέσο είναι (487,46, 496,54)». Με άλλα λόγια, η πρόταση

πιθανότητας «υπάρχει πιθανότητα 95% δηλ. $P(-t_{v,\alpha/2} < t^{*} < t_{v,\alpha/2}) = 95\%$, η αληθινή αλλά άγνωστη τιμή του μ να βρίσκεται στο διάστημα που εκτιμήσαμε» είναι σωστή μόνο πριν την εκτίμηση. Μετά την εκτίμηση ή θα είναι ή δεν θα είναι ο μέσος μ στο διάστημα που εκτιμήθηκε.

Παράδειγμα 8.2

Το ΔΣ μεγάλης επιχείρησης επιθυμεί να εκτιμήσει με βαθμό εμπιστοσύνης 90% το μέσο ύψος των αποδοχών των εργαζομένων. Για την εκτίμηση του μέσου ύψους των αποδοχών ο οικονομικός διευθυντής πήρε τυχαίο δείγμα 49 εκκαθαριστικών μισθοδοσίας από το λογιστήριο, από το οποίο υπολόγισε τιμή δειγματικού μέσου 1.400€ και διακύμανσης 250.000€. Ποια είναι η εκτίμηση του υπόψη διαστήματος εμπιστοσύνης;

Απάντηση:

Εφαρμόζουμε το κεντρικό οριακό θεώρημα επειδή έχουμε μεγάλο δείγμα n=49. Επομένως εφαρμόζοντας τα 4 στάδια της διαδικασίας του Πίνακα 8.1 για την εκτίμηση διαστήματος εμπιστοσύνης θα έχουμε:

Βήμα 1ο

Εκτιμήτρια. Η εκτιμήτρια της άγνωστης πληθυσμιακής παραμέτρου μ είναι $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ την οποία χρησιμοποίησε ο εν λόγω διευθυντής παραγωγής και πήρε ως σημειακή εκτίμηση $\bar{X}$=1.400. Το τυπικό της σφάλμα είναι $s_{\bar{X}} = \sqrt{\frac{250.000}{49}} = 71,429$.

Βήμα 2ο

Στατιστική ελέγχου, $Y = f(\hat{\theta}, \Theta)$. Η τυχαία μεταβλητή και μάλιστα του τυποποιημένου μέσου, που μας υποδεικνύει η κατανομή δειγματοληψίας στη συγκεκριμένη περίπτωση (άγνωστη πληθυσμιακή διακύμανση σ^2, αλλά μεγάλο δείγμα), δηλ. το κεντρικόοριακό θεώρημα, είναι συνάρτηση τόσο της εκτιμήτριας όσο και της παραμέτρου,

δηλ. $Z^{*}(\equiv \frac{\bar{X}-\mu}{s_{\bar{X}}}) \sim Z(0,1)$.

Βήμα 3ο

Επίπεδο εμπιστοσύνης, κρίσιμες τιμές. Το επιλεγμένο επίπεδο εμπιστοσύνης 1-α είναι 1-0,10=0,90 ή 90% και προσδιορίζει κριτικές τιμές, τις οποίες βρίσκουμε από τους Πίνακες της τυπικής κανονικής, τις $|z_{\alpha/2}| = |z_{0,05}| = 1,645$.

Βήμα 4ο

Όρια εμπιστοσύνης. Υπολογίζουμε τα όρια εμπιστοσύνης 90%, βάσει της πιθανότητας

P(l<Θ<u)=90% $P(|Z^*| < z_{\alpha/2}) = 90\%$ ή $P(-z_{\alpha/2} < Z^* < z_{\alpha/2}) = 90\%$

ή $P(\bar{X} - z_{\alpha/2} \cdot s_{\bar{X}} < \mu < \bar{X} + z_{\alpha/2} \cdot s_{\bar{X}}) = 90\%$. Επομένως το ζητούμενο 90% διάστημα εμπιστοσύνης θα είναι $\bar{X} \pm z_{\alpha/2} \cdot s_{\bar{X}} = 1.400 \pm 1,645 \cdot 71,429 = (1.282,51, 1.517,49)$

Άρα, το ζητούμενο από το ΔΣ διάστημα εμπιστοσύνης 90% είναι:
1.282,51 < μ < 1.517,49.

Αναφορικά με την εκτίμηση της πληθυσμιακής τυπικής απόκλισης σ , σε διάστημα, με βαθμό εμπιστοσύνης 1-α, προσδιορίζεται και αυτή από την κατανομή δειγματοληψίας της s. Οι δύο συνηθέστερες περιπτώσεις είναι:

- για γνωστή πληθυσμιακή διακύμανση σ^2 και τυχαία δειγματοληψία χωρίς επανάθεση ή από πεπερασμένο, κανονικό όμως, πληθυσμό, όπου η κατανομή της s είναι κανονική δηλ. $s \sim N(\sigma, \frac{\sigma^2}{2n})$, και έτσι χρησιμοποιούμε ως στατιστική ελέγχου την

$$Z^* (\equiv \frac{s-\sigma}{\sigma/\sqrt{2n}}) \sim Z(0,1).$$

- για άγνωστη πληθυσμιακή διακύμανση σ^2 και τυχαία δειγματοληψία με επανάθεση ή από άπειρο κανονικό πληθυσμό, όπου η κατανομή της s ακολουθεί την t-Student δηλ. $\gamma\lambda.\ s \sim t_\nu,\ \ \nu = n$, και γι' αυτό χρησιμοποιούμε ως στατιστική ελέγχου την

$$t^* (\equiv \frac{s-\sigma}{s/\sqrt{2n}}) \sim t_\nu.$$

Τα παραπάνω σχετικά με την κατανομή δειγματοληψίας της s τα εφαρμόζουμε στη μεθοδολογία της Κλασικής εκτίμησης σε διάστημα εμπιστοσύνης που συνοπτικά δίνεται στον Πίνακα 8.1 και παίρνουμε τις ζητούμενες εκτιμήσεις της πληθυσμιακής τυπικής απόκλισης σ.

Παράδειγμα 8.3

Ερευνητές του Τ.Ε.Ι. Κρήτης στην προσπάθειά τους να εκτιμήσουν τη μόλυνση του νερού σε χείμαρρο πλησίον βιομηχανικής περιοχής κατέγραψαν από 15 δείγματα σε ολόκληρο το μήκος του έτους (στρωματοποιημένη δυσανάλογη τυχαία δειγματοληψία) για συγκεκριμένο χαρακτηριστικό $\sum_i (x_i - \bar{X})^2 = 508,06$ mg. Να εκτιμηθεί 99% διάστημα εμπιστοσύνης της πληθυσμιακής τυπικής απόκλισης σ.

Απάντηση:

Εφαρμόζουμε τη γνωστή μεθοδολογία διαστημικής εκτίμησης, λαμβάνοντας υπόψη ότι το μικρό μέγεθος του δείγματος μας οδηγεί στην υιοθέτηση της t-Student θεωρητικής κατανομής για την πιθανοθεωρητική συμπεριφορά της δειγματικής τυπικής απόκλισης.

Βήμα 1ο

Εκτιμήτρια. Η εκτιμήτρια της άγνωστης πληθυσμιακής παραμέτρου σ είναι $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{X})^2}$ την οποία χρησιμοποιούμε για να πάρουμε ως σημειακή εκτίμηση $s=\sqrt{\frac{1}{15-1}508,06}=6,024$.

Το τυπικό της σφάλμα είναι $s_s=\frac{s}{\sqrt{2n}}=\frac{6,024}{\sqrt{30}}=1,10$.

Βήμα 2ο

Στατιστική ελέγχου, $Y=f(\hat{\theta},\Theta)$. Η τυχαία μεταβλητή και μάλιστα της τυποποιημένης τυπικής απόκλισης, που μας υποδεικνύει η κατανομή δειγματοληψίας στη συγκεκριμένη περίπτωση (άγνωστη πληθυσμιακή διακύμανση σ^2, και μικρό δείγμα) είναι συνάρτηση τόσο της εκτιμήτριας όσο και της παραμέτρου, δηλ. $t^{*}(\equiv\frac{s-\sigma}{s/\sqrt{2n}})\sim t_{\nu}$.

Βήμα 3ο

Επίπεδο εμπιστοσύνης, κρίσιμες τιμές. Το επιλεγμένο επίπεδο εμπιστοσύνης 1-α είναι 1-0,01=0,99 ή 99% και προσδιορίζει κριτικές τιμές, τις οποίες βρίσκουμε από τους Πίνακες της t-Student, στην $t_{15-1}=t_{14}$, τις $|t_{14,0,01}|=2,977$.

Βήμα 4ο

Όρια εμπιστοσύνης. Υπολογίζουμε τα όρια εμπιστοσύνης 99%, βάσει της πιθανότητας P(l<Θ<u)=99% $P(|t^{*}|<t_{\nu,\alpha/2})=99\%$ ή $P(-t_{\nu,\alpha/2}<t^{*}<t_{\nu,\alpha/2})=99\%$

ή $P(s-t_{\nu,\alpha/2}\cdot s_s<\sigma<s+t_{\nu,\alpha/2}\cdot s_s)=99\%$.

Επομένως το ζητούμενο 95% διάστημα εμπιστοσύνης θα είναι:
$s\pm t_{\nu,\alpha/2}\cdot s_s=6,024\pm2,977\cdot1,10=(2,750,9,298)$.

Άρα, το ζητούμενο με βαθμό εμπιστοσύνης 99% διάστημα της πληθυσμιακής τυπικής απόκλισης είναι 2,750 < σ < 9,298.

Για την εκτίμηση σε διάστημα εμπιστοσύνης του πληθυσμιακού ποσοστού P βασιζόμαστε και εδώ στην κατανομή δειγματοληψίας του δειγματικού $\hat{p}$, η οποία όπως εξηγήσαμε στο τμήμα 7.3.5 για μεγάλα δείγματα (n≥30) ακολουθεί τον κανονικό νόμο $\hat{p} \sim N(P, \frac{PQ}{n})$, και επομένως εκείνη του τυποποιημένου $\hat{p}$ την τυπική κανονική, με αποτέλεσμα να μπορούμε να χρησιμοποιήσουμε ως στατιστική ελέγχου την $Z^{*}(\equiv \frac{\hat{p}-P}{s_{\hat{p}}} = \frac{\hat{p}-P}{\sqrt{\frac{PQ}{n}}}) \sim Z(0,1)$.

Εφαρμόζοντας ανάλογα και εδώ τη μεθοδολογία εκτίμησης παραμέτρων σε διάστημα που δείξαμε στον Πίνακα 8.1 καταλήγουμε αβίαστα στις ζητούμενες εκτιμήσεις.

Παράδειγμα 8.4

Σε έρευνα αγοράς νέο-εισαγόμενου προϊόντος στην αγορά 60 καταναλωτές απάντησαν ότι θα μεταστραφούν προς αυτό γιατί τους ικανοποιεί η σχέση ποιότητα-τιμή. Το τυχαίο δείγμα του τμήματος marketing της υπόψη εταιρείας αφορούσε την πρωτεύουσα περιφέρειας και έγινε με το σχέδιο της συστηματικής δειγματοληψίας, ενώ είχε μέγεθος 150. Με αυτά τα δεδομένα μόνο, μπορείτε να εκτιμήσετε σε διάστημα 95% το άγνωστο πληθυσμιακό ποσοστό P για την επιδοκιμασία του νέου αυτού προϊόντος;

Απάντηση:

Δεδομένου ότι το τυχαίο δείγμα είναι μεγάλο και ο πληθυσμός διωνυμικός η κατανομή δειγματοληψίας του $\hat{p}$ είναι η κανονική, οπότε οι εκτιμήσεις μας θα γίνουν υπό την τυπική κανονική καμπύλη. Εφαρμόζουμε τη διαδικασία διαστημικής εκτίμησης του Πίνακα 8.1.

Βήμα 1ο

Εκτιμήτρια. Η εκτιμήτρια της άγνωστης πληθυσμιακής παραμέτρου P είναι $\hat{p} = \frac{x}{n}$ η την οποία χρησιμοποιούμε για τη σημειακή εκτίμηση $\hat{p} = \frac{x}{n} = \frac{60}{150} = 0,40$.

Το τυπικό της σφάλμα είναι $s_{\hat{p}} = \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} = \sqrt{\frac{0,40\cdot 0,60}{150}} = 0,04$.

Βήμα 2ο

Στατιστική ελέγχου, $Y = f(\hat{\theta}, \Theta)$. Η τυχαία μεταβλητή και μάλιστα του τυποποιημένου ποσοστού, που μας υποδεικνύει η κατανομή δειγματοληψίας στη συγκεκριμένη περίπτωση (άγνωστη πληθυσμιακή διακύμανση σ^2, αλλά μεγάλο δείγμα), είναι συνάρτηση τόσο της εκτιμήτριας όσο και της παραμέτρου,

δηλ. $Z^{*}(\equiv \frac{\hat{p}-P}{s_{\hat{p}}} = \frac{\hat{p}-P}{\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}}) \sim Z(0,1)$..

Βήμα 3ο

Επίπεδο εμπιστοσύνης, κρίσιμες τιμές. Το επιλεγμένο επίπεδο εμπιστοσύνης 1-α είναι 1-0,05=0,95 ή 95% και προσδιορίζει κριτικές τιμές, τις οποίες βρίσκουμε από τους Πίνακες της τυπικής κανονικής, τις $|z_{\alpha/2}|=|z_{0,025}|=1,96$.

Βήμα 4ο

Όρια εμπιστοσύνης. Υπολογίζουμε τα όρια εμπιστοσύνης 95%, βάσει της πιθανότητας P(l<Θ<u)=95% ή $P(|Z^*|<z_{\alpha/2})=95\%$ ή $P(-z_{\alpha/2}<Z^*<z_{\alpha/2})=95\%$ ή $P(\hat{p}-z_{\alpha/2}\cdot s_{\hat{p}}<P<\hat{p}+z_{\alpha/2}\cdot s_{\hat{p}})=95\%$. Επομένως το ζητούμενο 95% διάστημα εμπιστοσύνης θα είναι $\hat{p}\pm z_{\alpha/2}\cdot s_{\hat{p}}=0,40\pm1,96\cdot0,04=(0,322,\ 0,478)$

Άρα, το ζητούμενο από την εταιρεία διάστημα εμπιστοσύνης 95% είναι 32,2% < P < 47,8%, το οποίο θεωρείται ικανοποιητικό για το στάδιο ανάπτυξης του προϊόντος στην αγορά.

8.4 Έλεγχος στατιστικών υποθέσεων

8.4.1 Έννοια και μεθοδολογία κλασικού παραμετρικού ελέγχου υποθέσεων

Όταν λέμε έλεγχος υποθέσεων, εννοούμε τη στατιστική μεθοδολογία που ακολουθείται, προκειμένου να διατυπωθούν κρίσεις ή να ληφθούν αποφάσεις, κάτω από συνθήκες αβεβαιότητας, όταν χρησιμοποιώντας τη διαθέσιμη από τυχαίο δείγμα πληροφόρηση, σκοπεύουμε σε κρίση ή απόφαση αναφορικά με τη μορφή του δειγματοληπτούμενου πληθυσμού.

Οι συνθήκες αβεβαιότητας, οι οποίες μας υποχρεώνουν να εξετάσουμε πολύ μικρό τμήμα του πληθυσμού, το τυχαίο δείγμα, για να είναι αντιπροσωπευτικό της δομής του, μας αναγκάζουν, δεδομένου ότι πάντα θα υπάρχει δειγματοληπτικό λάθος $(\hat{\theta}-\Theta)$,, να στοχεύουμε μόνο στην ελαχιστοποίηση της πιθανότητας εμφάνισής του, και γενικότερα να μιλάμε σε όρους πιθανοτήτων. Έτσι είναι οι συνθήκες αβεβαιότητας που εξηγούν το επίθετο «στατιστικός» έλεγχος υποθέσεων ή έλεγχος στατιστικών υποθέσεων.

Η μορφή του πληθυσμού μπορεί να προσδιοριστεί αν γνωρίζουμε τις τιμές βασικών του παραμέτρων. Από εδώ πηγάζει ο όρος «παραμετρικός» έλεγχος στατιστικών υποθέσεων.

Επίσης, η μεθοδολογία επαγωγής των συμπερασμάτων του τυχαίου δείγματος στο γεννήτορα πληθυσμό του, προσδιορίζει το επίθετο «Κλασικός» παραμετρικός έλεγχος στατιστικών υποθέσεων.

Ο έλεγχος στατιστικών υποθέσεων είναι ίσως το σημαντικότερο τμήμα της επαγωγικής στατιστικής μεθοδολογίας, αφού υποδεικνύει τη λήψη συγκεκριμένων αποφάσεων. Η εκτίμηση μιας παραμέτρου, είτε σημειακή είτε σε διάστημα εμπιστοσύνης, δεν ολοκληρώνει

ποτέ την απόφαση μας για τον πληθυσμό. Κι αυτό ανεξάρτητα από την ποιότητα του εκτιμητή που χρησιμοποιήσαμε, που όπως είδαμε από τις επιθυμητές ιδιότητες π.χ. της αμεροληψίας, αποτελεσματικότητας επάρκειας και συνέπειας, μπορεί η μέθοδος εύρεσής του (πχ. ελαχίστων τετραγώνων) να τις εξασφαλίζει. Είναι δυνατόν όμως να μην υπάρχει δειγματοληπτικό σφάλμα $(\hat{\theta}-\Theta)$; Δηλ. διαφορά ανάμεσα στην εκτίμηση με βάση τα στοιχεία του δείγματος και στην αληθινή αλλά άγνωστη τιμή της παραμέτρου στον πληθυσμό; Η απάντηση είναι αρνητική, γιατί: Πρώτον, εξετάζουμε τυχαίο μεν, δείγμα όμως που είναι πολύ μικρό τμήμα του πληθυσμού, κι επομένως κάποιο χαρακτηριστικό του πληθυσμού να μην «αντιστοιχεί» πλήρως στη δομή του δείγματός μας. Δεύτερον, είναι δυνατόν να επιλέξαμε το τυχαίο δείγμα λάθος χρονική περίοδο, με την έννοια, την περίοδο της δειγματοληψίας μας να είχαν συμβεί εξαιρετικά γεγονότα που αλλοίωσαν τη συνήθη μορφή της κατανομής του πληθυσμού. Τρίτον, μπορεί οι εμπλεκόμενοι στη δειγματοληψία και μεταφορά των δεδομένων στους Η/Υ να έκαναν λάθη.

Αλλά ακόμα και στην ιδανική περίπτωση όπου έχουμε ελαχιστοποιήσει την πιθανότητα του δειγματοληπτικού σφάλματος, η απόκλιση αυτή σε απόλυτες τιμές, για παράδειγμα στην περίπτωση του μέσου $|\bar{X}-\mu|$, πρέπει να αξιολογηθεί στατιστικά πόσο σπουδαία είναι. Στη στατιστική θεωρία μάλιστα, το απόλυτο αυτό δειγματοληπτικό σφάλμα μετριέται σε όρους τυπικής απόκλισης της στατιστικής (δηλ. σε όρους τυπικού σφάλματος), έτσι ώστε να τυποποιηθεί $\frac{|\bar{X}-\mu|}{\sigma_{\bar{X}}}$ ή $\frac{|\bar{X}-\mu|}{s_{\bar{X}}}$ και να μπορεί να αξιολογηθεί σε όρους πιθανότητας, π.χ. κάτω από την κανονική καμπύλη Ζ(0,1).

Έτσι, το ζητούμενο εδώ, είναι να μετρήσουμε αν αυτό το δειγματοληπτικό σφάλμα (που θα υπάρχει πάντα για τους ενδεικτικούς τουλάχιστον λόγους που προαναφέραμε) είναι στατιστικά (δηλ. με μετρήσιμη πιθανότητα π.χ. 1-α) σημαντικό.

Η μεθοδολογία του ελέγχου στατιστικών υποθέσεων είναι ανάλογη εκείνης της εκτίμησης πληθυσμιακών παραμέτρων σε διάστημα εμπιστοσύνης.

Βήμα 1ο Διατύπωση Υποθέσεων. Κατ΄αρχάς, διατυπώνουμε μια πρόταση αναφορικά με το υπό εξέταση χαρακτηριστικό του πληθυσμού, που περιγράφεται με κάποια παράμετρο, έστω, το μέσο μ. Η πρόταση αυτή ονομάζεται στατιστική υπόθεση και φυσικά αφορά την πληθυσμιακή παράμετρο η οποία είναι βέβαια και ο σκοπός της στατιστικής μας συμπερασματολογίας. Η στατιστική υπόθεση είναι αυτή που πρέπει να αντιπαρατεθεί με τα δεδομένα του δείγματός μας για να μπορέσουμε να μετρήσουμε σε όρους πιθανότητας το προαναφερθέν τυποποιημένο δειγματοληπτικό σφάλμα.

Η πρόταση πιθανότητας που κάνουμε για την πληθυσμιακή παράμετρο, πρέπει να χωρίζει το δειγματικό της χώρο, δηλ. το σύνολο των δυνατών ενδεχομένων της, σε δύο αμοιβαία αποκλειόμενα ενδεχόμενα ή υποσύνολα, έτσι ώστε αν με κάποια επιστημονική μεθοδολογία οδηγηθούμε να αποκλείσουμε την πιθανότητα εμφάνισης του ενός να είμαστε αναγκασμένοι να μην απορρίψουμε το άλλο. Έτσι, οι στατιστικές υποθέσεις, ή προτάσεις πιθανότητας, τίθενται πάντα ανά ζεύγη, ακριβώς για να εξαντλούν το δειγματικό χώρο της άγνωστης αλλά αληθινής πληθυσμιακής παραμέτρου. Η πρώτη από το ζεύγος των υποθέσεων ονομάζεται βασική ή ελεγχόμενη ή μηδέν υπόθεση (null hypothesis) και συμβολίζεται H_0. Η βασική υπόθεση, εκφράζει το status quo ή την κατάσταση όπως πιστεύουμε ότι ισχύει μέχρι σήμερα. Η μηδενική υπόθεση όπως αλλιώς αποκαλείται η H_0, μεταφράζει το γεγονός ότι στους περισσότερους ελέγχους προσπαθούμε να αποφασίσουμε

αν πρόκειται για μηδενική διαφορά (όχι επίδραση) ανάμεσα στην καθ' υπόθεση γνωστή τιμή της άγνωστης πληθυσμιακής παραμέτρου και κάποιας που φέρεται ότι πρόκειται να λάβει τη θέση της. Η δεύτερη από το ζεύγος των στατιστικών υποθέσεων ονομάζεται εναλλακτική (alternative hypothesis) και συμβολίζεται H_1 ή H_A. Η εναλλακτική υπόθεση εκφράζει την τάση που δείχνουν τα εμπειρικά δεδομένα, ή οι ενδείξεις που έχουμε ότι το ισχύον status quo αλλάζει. Η εναλλακτική υπόθεση, έχει ιδιαίτερη βαρύτητα έναντι της βασικής, αφού είναι αυτή που εκφράζει την «πεποίθησή μας» για το αποτέλεσμα του ελέγχου. Η εναλλακτική μπορεί να διατυπωθεί με τρείς διαφορετικούς τρόπους οι οποίοι προσδιορίζονται κάθε φορά από τη φύση του προβλήματος που εξετάζουμε.

Η φύση των στατιστικών υποθέσεων, που ταυτοποιούν ή εξειδικεύουν το στατιστικό πρόβλημα για τη λύση του οποίου κάνουμε τον έλεγχο, μπορεί να γίνει σαφέστερη με τη χρήση του παραδείγματος.

Παράδειγμα 8.5

Από τυχαίο δείγμα 50 φοιτητών στο μάθημα της Στατιστικής Ι, σε κάποιο ακαδημαϊκό τμήμα, υπολογίστηκε μέσος 7/10. Τίθενται προς έλεγχο τα ερωτήματα **α)** είναι 7/10 η μέση επίδοση ή μήπως είναι διαφορετική; **β)** μήπως είναι σωστοί οι ισχυρισμοί ορισμένων φοιτητών ότι η μέση επίδοση είναι μεγαλύτερη από 7/10; ή **γ)** μήπως η άποψη ορισμένων μελών ΔΕΠ ότι η μέση επίδοση των φοιτητών είναι μικρότερη από 7/10 είναι η σωστή;

Απάντηση:

α) Για την πρώτη περίπτωση, ο έλεγχος λέγεται δίπλευρος, και οι στατιστικές υποθέσεις τίθενται ως εξής:

H_0: μ=7/10 ή $\bar{X}-\mu=0$

H_1: μ≠7/10 ή $\bar{X}-\mu\neq0$

Ο τρόπος που έχουν τεθεί οι δύο υποθέσεις εξασφαλίζει τη βασική πρόβλεψη μερισμού του δειγματικού χώρου του πληθυσμιακού μέσου σε δύο ξένα ενδεχόμενα. Έτσι αν απορριφθεί η βασική θα γίνει αποδεκτή η εναλλακτική χωρίς καμία αμφιβολία για το συμπέρασμα ή απόφαση που θα πάρουμε. Η εναλλακτική υπόθεση εδώ είναι μια σύνθετη υπόθεση με την έννοια ότι δεν παίρνει συγκεκριμένη αριθμητική τιμή αλλά σύνολο τιμών.

β) Για τη δεύτερη περίπτωση ο έλεγχος λέγεται μονόπλευρος πάνω, και οι στατιστικές υποθέσεις τίθενται ως εξής:

H_0: μ=7/10
H_1: μ>7/10

Εδώ όμως αν απορριφθεί η βασική, που είναι και απλή, γιατί εξειδικεύει πλήρως την τιμή της πληθυσμιακής παραμέτρου, η αποδοχή της εναλλακτικής δεν σημαίνει υποχρεωτικά ότι ο μέσος δεν είναι πλέον 7/10 γιατί θα μπορούσε να ήταν <7/10. Επομένως είναι προτιμότερο, για να ικανοποιείται πλήρως η αρχή του μερισμού του δειγματικού χώρου του πληθυσμιακού μέσου σε δύο ξένα ενδεχόμενα, να τεθούν ως εξής:

H_0: $\mu \leq 7/10$
H_1: $\mu > 7/10$

β) Για την τρίτη περίπτωση ο έλεγχος λέγεται μονόπλευρος κάτω, και οι στατιστικές υποθέσεις πρέπει να τεθούν ως εξής:

H_0: $\mu \geq 7/10$
H_1: $\mu < 7/10$

Βήμα 2ο Στατιστική Ελέγχου. Επειδή η διεξαγωγή του ελέγχου αφορά συμπεριφορές υπό συνθήκες αβεβαιότητας, δηλ. μετρήσεις πειραμάτων τύχης που καταγράφονται στις τιμές των τυχαίων μεταβλητών μας, είναι απολύτως απαραίτητο να γνωρίζουμε την πιθανοθεωρητική δομή του φαινομένου, δηλ. το θεωρητικό νόμο που ακολουθεί η εκτιμήτριά μας, για να πιθανολογήσουμε με τη βοήθειά του. Διαφορετικά δεν μπορούμε να μετρήσουμε τη διαφορά του δειγματοληπτικού σφάλματος σε όρους πιθανότητας και επομένως, δεν μπορούμε να πούμε αν είναι στατιστικά σημαντική π.χ. η διαφορά $\bar{X}-\mu$. Επομένως έχουμε ανάγκη, όπως και στην εκτιμητική, της κατάλληλης στατιστικής ελέγχου η πιθανοθεωρητική συμπεριφορά της οποίας, συνιστά το εργαλείο με βάση το οποίο θα αποφασίσουμε στατιστικά, δηλ. σε όρους πιθανότητας. Για παράδειγμα, στο προηγούμενο παράδειγμα ελέγχου της μέσης πληθυσμιακής επίδοσης των φοιτητών στη Στατιστική Ι, κάτω από το Κεντρικό Οριακό Θεώρημα, εξαιτίας του μεγάλου δείγματος n=50 θα μπορούσε να χρησιμοποιηθεί η στατιστική ελέγχου $Z^{*}(\equiv \frac{\bar{X}-\mu}{s_{\bar{x}}}) \sim Z(0,1)$. Έτσι, μπορούμε να υπολογίσουμε πιθανότητες υπό την τυπική κανονική και να μετατρέψουμε τις συνθήκες αβεβαιότητας σε συνθήκες κινδύνου αφού πάντα υπάρχει το επίπεδο σημαντικότητας α που εκφράζει μια πιθανότητα λάθους απόφασης.

Βήμα 3ο Κριτήριο Ελέγχου ή Κανόνας Απόφασης. Ακόμα και η επιλογή της κατάλληλης για την κάθε περίπτωση στατιστικής ελέγχου, δεν αρκεί (είναι αναγκαία αλλά όχι ικανή συνθήκη) για να αποφασίσουμε για το στατιστικό πρόβλημα της τιμής ή συνόλου τιμών στο οποίο κυμαίνεται κάποια πληθυσμιακή παράμετρος. Πρέπει να θέσουμε και κάποιο κριτήριο με βάση το οποίο θα παίρνουμε τις αποφάσεις. Το κριτήριο αυτό έχει να κάνει με τα όρια που θα χωρίζουν το δειγματικό χώρο της παραμέτρου, ο οποίος μετριέται υπό την καμπύλη της στατιστικής ελέγχου (βλ. Βήμα 3ο διαστημικής εκτίμησης), σε δύο ξένα υποσύνολα ή αμοιβαία αποκλειόμενες περιοχές, την περιοχή αποδοχής της H_0 και την περιοχή απόρριψής της. Το κριτήριο αυτό στην κλασική παραμετρική επαγωγική είναι η μεγιστοποίηση της δύναμης του κριτηρίου που δίνεται από την πιθανότητα:

$P(\text{Απόρριψης } H_0 \mid H_0 \text{ λάθος}) = 1-\beta$	**(8.13)**

Για να αντιληφθεί σε αδρές γραμμές ο αναγνώστης την έννοια αυτή πρέπει να τοποθετήσουμε τη φύση του προβλήματος του ελέγχου στατιστικών υποθέσεων ή της διαδικασίας στατιστικής απόφασης, στο δειγματικό της χώρο, ο οποίος εδώ παρουσιάζεται σε μορφή πίνακα συνάφειας όπως παρακάτω.

Πίνακας 8.2 Δυνατά αποτελέσματα Ελέγχου Υποθέσεων

	Πραγματική κατάσταση	
Απόφαση	H_0 αληθινή	H_0 λανθασμένη
Δεν απορρίπτεται η H_0	**Σωστή απόφαση Πιθανότητα 1 – α**	Σφάλμα Τύπου ΙΙ Πιθανότητα β
Απορρίπτεται η H_0	Σφάλμα Τύπου Ι Πιθανότητα α	**Σωστή απόφαση Πιθανότητα 1 – β**

Από την εξέταση του Πίνακα 8.2, πρέπει τώρα να ξεκαθαρίζει η έννοια του επιπέδου ή βαθμού σημαντικότητας α που αναφέραμε στη μεθοδολογία της εκτιμητικής σε διάστημα, και έπαιζε καθοριστικό ρόλο στον ορισμό των κρίσιμων τιμών της στατιστικής ελέγχου, οι οποίες προσδιόριζαν, σε συνδυασμό με το τυπικό σφάλμα της εκτιμήτριας αυτής, τα όρια του ζητούμενου διαστήματος με πιθανότητα εμπιστοσύνης 1-α (κελί 1,1 του πίνακα 8.2). Έτσι, μια μορφή σφάλματος στον έλεγχο ή λάθους στατιστικής απόφασης είναι το σφάλμα τύπου Ι.

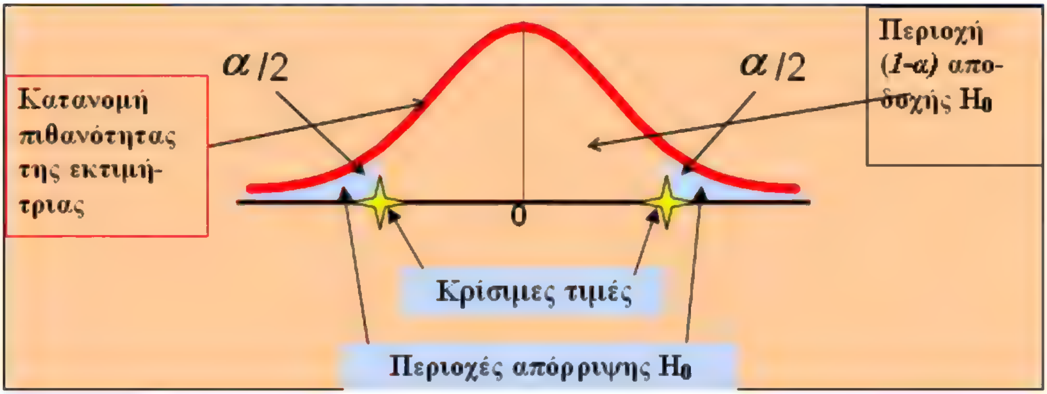
$$P(\text{Απόρριψης } H_0 \mid H_0 \text{ σωστή}) = \alpha \qquad \textbf{(8.14)}$$

Στο σφάλμα τύπου ΙΙ ή πιθανότητα β η απόφασή μας είναι η αντίθετη απ' ότι προηγουμένως στην (8.14), δηλ.

$$P(\text{μη} - \text{απόρριψης } H_0 \mid H_0 \text{ λάθος}) = \beta \qquad \textbf{(8.15)}$$

Με βάση λοιπόν το δειγματικό χώρο της πληθυσμιακής παραμέτρου που παραστατικά δίνεται στον Πίνακα 8.2 , το κριτήριο ελέγχου ή κανόνας απόφασης τον μερίζει στην περιοχή αποδοχής της H_0 με πιθανότητα εμπιστοσύνης 1-α και στην περιοχή απόρριψης της H_0 με βαθμό σημαντικότητας α. Τα παραπάνω με αναφορά τον δίπλευρο έλεγχο του μέσου φαίνονται παραστατικά στο Σχήμα 8.2.

Σχήμα 8.2 Κανόνας απόφασης για το δίπλευρο έλεγχο του μέσου



Είναι πλέον σαφές ότι το κριτήριο ελέγχου ή κανόνας απόφασης εξαρτάται από τον τρόπο διατύπωσης ή εξειδίκευσης των στατιστικών υποθέσεων. Έτσι, με αναφορά στον ευρύτερα χρησιμοποιούμενο έλεγχο υποθέσεων του μέσου (μεγάλο δείγμα, άγνωστη διακύμανση και επομένως χρήση του Κεντρικού Οριακού Θεωρήματος) τα κριτήρια ελέγχου διακρίνονται:

α) στο δίπλευρο έλεγχο π.χ. H_0: μ=7/10 έναντι της H_1: μ≠7/10
το κριτήριο είναι:

«Απορρίπτεται η H_0, εάν η τιμή της στατιστικής ελέγχου (εδώ Z*) από τις παρατηρήσεις του δείγματος «εντοπίζεται», έξω από τα όρια $|z_{a/2}|$ στις ουρές της κατανομής».

β) στο μονόπλευρο πάνω έλεγχο π.χ. H_0: μ ≤ 7/10 έναντι της H_1: μ > 7/10
το κριτήριο είναι:

«Απορρίπτεται η H_0 , εάν η τιμή της στατιστικής ελέγχου (εδώ Z*) από τις παρατηρήσεις του δείγματος «εντοπίζεται», πάνω από το πάνω όριο z_a στη δεξιά ουρά της κατανομής».

γ) στο μονόπλευρο κάτω έλεγχο π.χ. H_0: μ ≥ 7/10 έναντι της H_1: μ < 7/10
το κριτήριο είναι:

«Απορρίπτεται η H_0, εάν η τιμή της στατιστικής ελέγχου (εδώ Z*) από τις παρατηρήσεις του δείγματος «εντοπίζεται», κάτω από το κάτω όριο $-z_a$ στην αριστερή ουρά της κατανομής».

Τα παραπάνω κριτήρια για ελέγχους στατιστικών υποθέσεων υπό την τυπική κανονική δίνονται συνοπτικά στον παρακάτω Πίνακα.

Πίνακας 8.3 Κριτήρια ελέγχου με την τυπική κανονική κατανομή πιθανότητας f(z)

Δίπλευρος έλεγχος: Απορρίπτεται η H_0 εάν $Z^* > z_{\alpha/2}$ ή $Z^* < -z_{\alpha/2}$	**(8.16α)**
Μονόπλευρος πάνω έλεγχος: Απορρίπτεται η H_0 εάν $Z^* > z_\alpha$	**(8.16β)**
Μονόπλευρος κάτω έλεγχος: Απορρίπτεται η H_0 εάν $Z^* < -z_\alpha$	**(8.16γ)**

Τα αντίστοιχα κριτήρια για ελέγχους με την t-Student f(tv) είναι:

Πίνακας 8.4 Κριτήρια ελέγχου με την t-Student f(tv)

Δίπλευρος έλεγχος: Απορρίπτεται η H_0 εάν $t_v^* > t_{v,\alpha/2}$ ή $t_v^* < -t_{v,\alpha/2}$	**(8.17α)**
Μονόπλευρος πάνω έλεγχος: Απορρίπτεται η H_0 εάν $t_v^* > t_{v,\alpha}$	**(8.17β)**
Μονόπλευρος κάτω έλεγχος: Απορρίπτεται η H_0 εάν $t_v^* < -t_{v,\alpha}$	**(8.17γ)**

Τα αντίστοιχα κριτήρια για ελέγχους (π.χ. διακύμανσης) με την f($\chi^2_{\nu^*}$)είναι:

Πίνακας 8.5 Κριτήρια ελέγχου με την f($\chi^2_{\nu^*}$)

Δίπλευρος έλεγχος: Απορρίπτεται η H_0 εάν $\chi^2_{\nu^*} > \chi^2_{\nu,1-\alpha/2}$ ή $\chi^2_{\nu^*} < \chi^2_{\nu,\alpha/2}$	**(8.18α)**
Μονόπλευρος πάνω έλεγχος: Απορρίπτεται η H_0 εάν $\chi^2_{\nu^*} > \chi^2_{\nu,1-\alpha}$	**(8.18β)**
Μονόπλευρος κάτω έλεγχος: Απορρίπτεται η H_0 εάν $\chi^2_{\nu^*} < \chi^2_{\nu,\alpha}$	**(8.18γ)**

Τα αντίστοιχα κριτήρια για ελέγχους (π.χ. διαφοράς 2 διακυμάνσεων) με την f($F_{\nu,\nu}$) είναι:

Πίνακας 8.6 Κριτήρια ελέγχου με την f($F_{\nu,\nu}$)

Δίπλευρος έλεγχος: Απορρίπτεται η H_0 εάν $F^*_{\nu,\nu} > F_{\nu,\nu,1-\alpha/2}$ ή $F^*_{\nu,\nu} < F_{\nu,\nu,\alpha/2}$	**(8.19α)**
Μονόπλευρος πάνω έλεγχος: Απορρίπτεται η H_0 εάν $F^*_{\nu,\nu} > F_{\nu,\nu,1-\alpha}$	**(8.19β)**
Μονόπλευρος κάτω έλεγχος: Απορρίπτεται η H_0 εάν $F^*_{\nu_1,\nu_2} < F_{\nu_1,\nu_2,\alpha}$	**(8.19γ)**

Όπου $F^*_{\nu,\nu}$ χρησιμοποιείται ο μεγαλύτερος λόγος διακυμάνσεων.

Εναλλακτικά, αλλά ευρέως χρησιμοποιούμενος κανόνας απόφασης είναι το λεγόμενο ακριβές επίπεδο σημαντικότητας (exact level of significance) ή κρίσιμο μέγεθος του ελέγχου ή τιμή πιθανότητας (p-value). Η πιθανότητα αυτή εκφράζει το χαμηλότερο επίπεδο σημαντικότητας στο οποίο μπορεί να απορριφθεί η H_0, ή την πιθανότητα να πάρουμε τιμή της στατιστικής ελέγχου μεγαλύτερη ή ίση από την παρατηρούμενη δειγματική, δεδομένου ότι η H_0 είναι σωστή, συμβολικά μπορούμε να γράψουμε:

π.χ. στο δίπλευρο έλεγχο του μέσου $p-value \equiv P[(Z^{*} < -z_{a/2} \text{ ή } Z^{*} > z_{a/2}) \mid H_0 \text{ σωστή}]$	**(8.20)**
Κανόνας απόφασης: Απορρίπτεται η H_0 εάν $p-value < \alpha$	**(8.21)**

Βήμα 4ο Υπολογισμοί. Στο στάδιο αυτό αφενός, εφαρμόζουμε τα δεδομένα του τυχαίου δείγματος στη στατιστική ελέγχου ή εκτιμήτρια (π.χ. Z* στον έλεγχο μέσου, ΚΟΘ [10]) για να βρούμε πού εντοπίζεται η τυχαία μας μεταβλητή με βάση τις εμπειρικές ενδείξεις (στοιχεία δείγματος), και αφετέρου, με βάση το επιλεγμένο επίπεδο σημαντικότητας υπολογίζουμε τις κρίσιμες τιμές, που χωρίζουν το δειγματικό χώρο της υπό έλεγχο παραμέτρου στις αμοιβαία αποκλειόμενες περιοχές αποδοχής της H_0 με πιθανότητα εμπιστοσύνης 1-α και σε εκείνη της απόρριψης της H_0 με πιθανότητα α σφάλματος τύπου Ι .

Βήμα 5ο Στατιστικό Συμπέρασμα-Απόφαση. Από τη σύγκριση της τιμής της εκτιμήτριας ή στατιστικής ελέγχου (π.χ. Z* στον έλεγχο μέσου, με ΚΟΘ) με τις κρίσιμες τιμές που μερίζουν το δειγματικό χώρο στις περιοχές απόρριψης και αποδοχής της βασικής υπόθεσης, λαμβάνεται η στατιστική απόφαση, σύμφωνα με τα κριτήρια που δίνονται στους Πίνακες 8.3-8.6 ανάλογα την περίπτωση.

[10] ΚΟΘ Κεντρικό Οριακό Θεώρημα.

Συνοψίζοντας, η μεθοδολογία του κλασικού παραμετρικού ελέγχου στατιστικών υποθέσεων ακολουθεί τα βήματα που δίνονται στον παρακάτω πίνακα.

Πίνακας 8.7 Μεθοδολογία κλασικού παραμετρικού ελέγχου στατιστικών υποθέσεων

Βήμα 1ο	**Διατύπωση στατιστικών υποθέσεων**. Πολύ σημαντικό στάδιο γιατί εξειδικεύει το πρόβλημα του ελέγχου. Τρείς εναλλακτικές διατυπώσεις μερισμού του δειγματικού χώρου της ελεγχόμενης παραμέτρου, έστω μέσου μ Δίπλευρος H_0: μ=μ0 έναντι της H_1: μ≠μ0 Μονόπλευρος πάνω H_0: μ≤μ0 έναντι της H_1: μ>μ0 Μονόπλευρος κάτω H_0: μ≥μ0 έναντι της H_1: μ<μ0
Βήμα 2ο	**Στατιστική ελέγχου**, έστω, $Y = f(\hat{\theta}, \Theta)$. Επιλέγουμε τυχαία μεταβλητή, η οποία να είναι συνάρτηση τόσο της εκτιμήτριας όσο και της παραμέτρου, δηλ. $Y = f(\bar{X}, \mu)$, την κατανομή πιθανότητας της οποίας γνωρίζουμε από τους θεωρητικούς νόμους έτσι ώστε να μπορούμε να κάνουμε προτάσεις πιθανότητας υπό την καμπύλη της. Ανάλογα με την κατανομή δειγματοληψίας, αυτή μπορεί να είναι: π.χ. $Z^{*}(\equiv \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}) \sim Z(0,1)$.
Βήμα 3ο	**Κριτήριο ελέγχου ή κανόνας απόφασης**. Ανάλογα με τη διατύπωση των υποθέσεων, αν δηλ. έχουμε δικατάληκτο, μονοκατάληκτο κάτω ή πάνω κριτήρια, οι κανόνες απόφασης δίνονται στους Πίνακες 8.3-8.6 σύμφωνα με τις κατανομές των στατιστικών ελέγχου. Εναλλακτικά απορρίπτουμε τη H_0 εάν (p-value)< α.
Βήμα 4ο	**Υπολογισμοί**. α) Υπολογίζουμε τις κριτικές τιμές της εκτιμήτριας για το επιλεγόμενο επίπεδο σημαντικότητας α (π.χ. για το μέσο με ΚΟΘ είναι |zα/2|). Αυτές χωρίζουν το δειγματικό χώρο της ελεγχόμενης παραμέτρου στις περιοχές αποδοχής και απόρριψης της H_0.β) Εφαρμόζουμε τα δειγματικά δεδομένα στη στατιστική ελέγχου για να υπολογίσουμε την τιμή της διερευνώντας σε ποια από τις υπόψη δύο περιοχές εντοπίζεται.
Βήμα 5ο	**Στατιστικό συμπέρασμα-απόφαση**. Ανάλογα πού εντοπίζεται η τιμή της εκτιμήτριας από τα δειγματικά δεδομένα και με βάση το κριτήριο ελέγχου ή κανόνα απόφασης, στο επιλεγμένο επίπεδο σημαντικότητας, απορρίπτουμε ή δεν απορρίπτουμε τη βασική υπόθεση.

8.4.2 Εφαρμογές: Έλεγχοι υποθέσεων για το μέσο, την τυπική απόκλιση και την αναλογία

Παράδειγμα 8.6

Μέλη του ΔΣ ασφαλιστικής εταιρείας τα οποία πιστεύουν ότι οι αποζημιώσεις για συγκεκριμένη κατηγορία ασφαλειών είναι υψηλές και κατά μέσο όρο €600 ετησίως, ζήτησαν από τον οικονομικό διευθυντή να ελέγξει αυτήν την εικασία. Για τον έλεγχο της υπόθεσης αυτής ο οικονομικός διευθυντής επέλεξε τυχαίο δείγμα 36 τιμολογίων από όπου βρήκε μέσο €630 με διακύμανση 13.225€. Ποια πρέπει να ήταν η απάντηση του οικονομικού διευθυντή στο ΔΣ, εάν χρησιμοποίησε στατιστική συμπερασματολογία, και εάν ήθελε να είναι πολύ αυστηρός (γι' αυτό επέλεξε επίπεδο σημαντικότητας α=1%);

Απάντηση:

Ο έλεγχος γενικά όπως είδαμε παραπάνω, συνίσταται στην εξειδίκευση της στατιστικής ελέγχου, σχετικής με τη βασική υπόθεση H_0, καθώς επίσης και τον κανόνα απόφασης. Ως γνωστόν, τα δύο συστατικά του ελέγχου εξαρτώνται από την κατανομή δειγματοληψίας της υπόψη εκτιμήτριας.

Εδώ, εξαιτίας του μεγάλου μεγέθους του δείγματος n=36, παρά το γεγονός ότι δεν γνωρίζουμε τη διακύμανση του πληθυσμού, χρησιμοποιούμε το κεντρικό οριακό θεώρημα (ΚΟΘ), το οποίο προβλέπει κανονική κατανομή του δειγματικού μέσου και επομένως, επιλέγουμε στατιστική ελέγχου τον τυποποιημένο μέσο Z*. Με βάση τη μεθοδολογία που δείξαμε παραπάνω η απόφασή του οικονομικού διευθυντή ακολουθεί τα παρακάτω βήματα μέχρι την ολοκλήρωσή της.

Βήμα 1ο Διατύπωση Υποθέσεων

H_0: μ=600
H_1: μ≠600

Ο δίπλευρος έλεγχος δικαιολογείται από την ένδειξη από το δείγμα της μικρής απόκλισης της δειγματικής τιμής από την καθ' υπόθεση γνωστή πληθυσμιακή $|\bar{X}-\mu_0|$ (μ_0 καθ' υπόθεση γνωστός πληθυσμιακός).

Βήμα 2ο Επιλογή στατιστικής ελέγχου

$$Z^*(\equiv \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}) \sim Z(0,1)$$

Όπως προαναφέραμε εξαιτίας του μεγάλου δείγματος μπορούμε να χρησιμοποιήσουμε το ΚΟΘ, αντικαθιστώντας την άγνωστη πληθυσμιακή διακύμανση στον υπολογισμό του τυπικού σφάλματος, με την αμερόληπτη εκτίμησή της από το δείγμα.

Βήμα 3ο Κριτήριο ελέγχου ή κανόνας απόφασης

Κάτω από τις συνθήκες του τυχαίου αυτού πειράματος ο κανόνας απόφασης είναι «απορρίπτεται η H_0: μ=600 εάν η στατιστική ελέγχου με τα δειγματικά δεδομένα εντοπιστεί σε οποιοδήποτε των δύο άκρων της Z με όρια τα $z_{\alpha/2}$=z0,005», ή συμβολικά:

«Απορρίπτεται η H_0: μ=600 αν $Z^* > z_{\alpha/2}$ ή $Z^* < -z_{\alpha/2}$»

Εναλλακτικά,
«απορρίπτεται η H_0: μ=600 αν (p-value)< α »

Βήμα 4ο Υπολογισμοί

$$Z^* = \frac{\bar{X}-\mu}{s_{\bar{x}}} = \frac{630-600}{115/\sqrt{36}} = 1,565 \qquad |z_{\alpha/2}| = |z_{0,005}| = 2,576$$

$$p-value = P(\bar{X} > 630) = P(\frac{\bar{X}-\mu}{s_{\bar{x}}} > \frac{630-600}{115/6}) = P(Z > 1,565) = 0,0588$$

Βήμα 5ο Στατιστικό συμπέρασμα-απόφαση

Με βάση τον κανόνα απόφασης «Απορρίπτεται η H_0: μ=600 αν $Z^* > z_{\alpha/2}$ ή $Z^* < -z_{\alpha/2}$ » βρήκαμε Z*=1,565 ενώ από τους πίνακες της τυπικής έχουμε $|z_{\alpha/2}| = |z_{0,005}| = 2,576$ για α=1%. Επομένως, επειδή $Z^* < z_{\alpha/2}$ δεν απορρίπτουμε τη H_0: μ=600 αποδεχόμενοι ότι η διαφορά $|\bar{X}-\mu_0|$ δεν είναι στατιστικά σημαντική και οφείλεται στις τυχαίες κυμάνσεις της δειγματοληψίας. Εντούτοις, πάντα διατηρούμε επιφύλαξη, να έχουμε κάνει κάποιο δειγματοληπτικό λάθος.

Εναλλακτικά, με τον κανόνα απόφασης «απορρίπτεται η H_0: μ=600 αν (p-value)< α » επίσης καταλήγουμε στο ίδιο συμπέρασμα να μην απορρίψουμε τη βασική αφού p-value (=5,88%) > α (=1%). Με άλλα λόγια, για να απορρίψουμε την H_0: μ=600 θα έπρεπε να είχαμε θέσει επίπεδο σημαντικότητας π.χ. α>5,9% κάτι το οποίο θεωρείται υψηλό.

Παράδειγμα 8.7

Ομάδα ερευνητών ακαδημαϊκού τμήματος για τη διαδικασία της εσωτερικής αξιολόγησής του, επιδιώκει να ελέγξει αν ο μέσος όρος των φοιτητών που παρακολουθούν τα μαθήματα έχει αυξηθεί από τον προπενταετίας που έως 50. Για το λόγο αυτό επέλεξε τυχαίο δείγμα 25 μαθημάτων στα οποία υπολόγισε, κατά το τρέχον χειμερινό εξάμηνο, κατά μέσο όρο 53 παρόντες φοιτητές, με τυπική απόκλιση 10. Εάν οι ερευνητές έθεσαν βαθμό εμπιστοσύνης των εκτιμήσεων τους 90%, να δείξετε αν οι εμπειρικές ενδείξεις συμφωνούν με την προπενταετίας διαπίστωση.

Απάντηση:

Επειδή το δείγμα είναι μικρό (n<30), αφενός, υποχρεωτικά θα εργαστούμε κάτω από το νόμο του Gosset ή την t-Student κατανομή, και αφετέρου πρέπει να είμαστε περισσότερο αυστηροί στη διατύπωση των υποθέσεων, δηλ. έστω και μικρή ένδειξη διαφοράς $|\bar{X}-\mu_0|$ πρέπει να μας προβληματίζει σημαντικά προς την κατεύθυνση μονόπλευρων ελέγχων.

Βήμα 1ο Διατύπωση Υποθέσεων

H_0: μ≤50
H_1: μ>50

Βήμα 2ο Επιλογή στατιστικής ελέγχου

$$t_v^{\bullet}(=\frac{\bar{X}-\mu}{s_{\bar{X}}})>t_{v,a}$$

Βήμα 3ο Κριτήριο ελέγχου ή κανόνας απόφασης

Κάτω από τις συνθήκες του τυχαίου αυτού πειράματος ο κανόνας απόφασης είναι «απορρίπτεται η H_0: μ≤50 εάν η στατιστική ελέγχου με τα δειγματικά δεδομένα εντοπιστεί εντός της δεξιάς ουράς της $t_{v,a}$ με όριο το $t_{v,a}= t_{24,10\%}$», ή συμβολικά:

«Απορρίπτεται η H_0: μ≤50 αν $t_v^{\bullet}>t_{24,0.10}$»

Εναλλακτικά,
«απορρίπτεται η H_0: μ≤50 αν (p-value)< α »

Βήμα 4ο Υπολογισμοί

$$t_v^{\bullet}=\frac{\bar{X}-\mu}{s_{\bar{X}}}=\frac{53-50}{10/\sqrt{25}}=1,50 \qquad t_{24,0.10}=1,318$$

$$p-value=P(\bar{X}>53)=P(\frac{\bar{X}-\mu}{s_{\bar{X}}}>\frac{53-50}{10/5})=P(t_{24}>1,50)=0,0733$$

Βήμα 5ο Στατιστικό συμπέρασμα-απόφαση

Με βάση τον κανόνα απόφασης «Απορρίπτεται η H_0: μ≤50 αν $t_v^{\bullet}>t_{24,0.10}$ » βρήκαμε $t^{\bullet}$ =1,5 ενώ από τους πίνακες της t-Student για να έχουμε στη δεξιά ουρά της $t_{24,0,1}$ 10% βρίσκουμε κρίσιμη τιμή $t_{24,0,1}$=1,318(για α=10%).

Επομένως, επειδή $t^{\bullet}$ > $t_{24,0,1}$ απορρίπτουμε την H_0: μ≤50 αποδεχόμενοι ότι πρέπει να έχει αυξηθεί ο μέσος όρος παρακολούθησης μαθημάτων από τους φοιτητές (μ>50). Με άλλα λόγια, η διαφορά $|\bar{X}-\mu_0|$ είναι στατιστικά σημαντική και δεν μπορεί να αποδοθεί στις τυχαίες κυμάνσεις της δειγματοληψίας. Εντούτοις, πάντα διατηρούμε επιφύλαξη, να έχουμε κάνει κάποιο δειγματοληπτικό λάθος.

Εναλλακτικά, με τον κανόνα απόφασης «απορρίπτεται η H_0: μ≤50 αν (p-value)< α» επίσης καταλήγουμε στο ίδιο συμπέρασμα να απορρίψουμε τη βασική αφού p-value (=7,33%) < α (=10%).

Παράδειγμα 8.8

Επενδυτής, αυτοαποκαλούμενος ριψοκίνδυνος, είχε πέρυσι διακύμανση των αποδόσεων του χαρτοφυλακίου (Χ/Φ) των 15 μετοχών του 25% ενώ την ίδια περίοδο η αντίστοιχη διακύμανση του γενικού δείκτη ήταν 18%. Δικαιολογείται ο χαρακτηρισμός του επενδυτή ως ριψοκίνδυνου με βαθμό εμπιστοσύνης της εκτίμησής σας 95%;

Απάντηση:

Πρόκειται για έλεγχο διακύμανσης σ^2 όπου ως καθ' υπόθεση γνωστή πληθυσμιακή δεχόμαστε την σ_0^2 =0,18.

Βήμα 1ο Διατύπωση Υποθέσεων

H_0: $\sigma^2 = 0{,}18$
H_1: $\sigma^2 > 0{,}18$

Βήμα 2ο Επιλογή στατιστικής ελέγχου

Επειδή έχουμε μικρό δείγμα n=15 η κατανομή δειγματοληψίας της s^2 αποδεικνύεται ότι είναι ακολουθεί τη χ^2_v, όπου v=n-1, επομένως η στατιστική ελέγχου θα είναι:

$$\chi^2_{v^*} \left(= \frac{(n-1)\cdot s^2}{\sigma_0^2}\right) \sim \chi^2_v$$

Βήμα 3ο Κριτήριο ελέγχου ή κανόνας απόφασης

Κάτω από τις συνθήκες του τυχαίου αυτού πειράματος ο κανόνας απόφασης είναι «απορρίπτεται η H_0: $\sigma^2 = 0{,}18$ εάν η στατιστική ελέγχου με τα δειγματικά δεδομένα εντοπιστεί εντός της δεξιάς ουράς της $\chi^2_{v,\alpha}$ με όριο το $\chi^2_{v,\alpha} = \chi^2_{14,0{,}05}$», συμβολικά:

«Απορρίπτεται η H_0: $\sigma^2 = 0{,}18$ αν $\chi^2_{v^*} = \chi^2_{14,0{,}05}$»

Εναλλακτικά,
«απορρίπτεται η H_0: $\sigma^2 = 0{,}18$ αν (p-value)< α »

Βήμα 4ο Υπολογισμοί

$$\chi^2_{v^*} = \frac{(n-1)\cdot s^2}{\sigma_0^2} = \frac{(15-1)\cdot 0{,}25}{0{,}18} = 19{,}444 \qquad \chi^2_{v,\alpha} = \chi^2_{14,0{,}05} = 23{,}685$$

$$p-value = P(s^2 > 0{,}25) = P\left(\frac{(n-1)\cdot s^2}{\sigma_0^2} > \frac{(15-1)\cdot 0{,}25}{0{,}18}\right) = P(\chi^2_{14} > 19{,}444) = 0{,}1487$$

Βήμα 5ο Στατιστικό συμπέρασμα-απόφαση

Με βάση τον κανόνα απόφασης «Απορρίπτεται η H_0: σ^2 = 0,18 αν $\chi^2_{v^*} = \chi^2_{14,0,05}$». Έτσι, αφού $\chi^2_{v^*} = 19,444 < \chi^2_{14,0,05} = 23,685$ δεν απορρίπτουμε τη βασική υπόθεση.

Επομένως, το Χ/Φ του υπόψη επενδυτή πρέπει να θεωρηθεί ότι έχει διαφορετική διακύμανση από εκείνο του γενικού δείκτη, και επειδή έχουμε ένδειξη υψηλότερης διασποράς δεν μπορούμε να απορρίψουμε τον ισχυρισμό του ότι είναι ριψοκίνδυνος. Εντούτοις, πάντα διατηρούμε επιφύλαξη, να έχουμε κάνει κάποιο δειγματοληπτικό λάθος.

Εναλλακτικά, με τον κανόνα απόφασης «απορρίπτεται η H0: σ^2 = 0,18 αν (p-value)< α » επίσης καταλήγουμε στο ίδιο συμπέρασμα να μην απορρίψουμε τη βασική, αφού p-value (=14,87%) > α (=5%).

Παράδειγμα 8.9

Οικονομολόγοι του Ελεγκτικού Συνεδρίου γνώριζαν από την εμπειρία τους ότι τουλάχιστον 20% των τιμολογίων δαπανών ΟΤΑ περιλαμβάνουν σοβαρές υπερβάσεις των επιτρεπόμενων ορίων. Φέτος όμως, σε τυχαίο δείγμα 49 μεγάλων ΟΤΑ καταλόγισαν παρόμοιες παραβάσεις σε ποσοστό 32%. Επιτρέπεται να συμπεράνουμε, με βαθμό εμπιστοσύνης 10%, ότι αυτή η κατάσταση στους ΟΤΑ χειροτέρεψε;

Απάντηση:
Μας ενδιαφέρει να ελέγξουμε αν οι εμπειρικές μας ενδείξεις συνηγορούν υπέρ της αύξησης του θεωρούμενο ως «ισχύοντος» πληθυσμιακού ποσοστού P παραβάσεων 20%. Εξαιτίας του μεγάλου μεγέθους του δείγματος η κατανομή της εκτιμήτριας είναι η κανονική.

Βήμα 1ο Διατύπωση Υποθέσεων

H_0: P≤0,20
H_1: P>0,20

Βήμα 2ο Επιλογή στατιστικής ελέγχου

$$Z^*(\equiv \frac{\hat{p}-P}{s_{\hat{p}}}) \sim Z(0,1)$$

Βήμα 3ο Κριτήριο ελέγχου ή κανόνας απόφασης

«Απορρίπτεται η H_0: P≤0,20 αν $Z^* > z_\alpha$».

Εναλλακτικά, «απορρίπτεται η H_0: P≤0,20 αν (p-value) < α »

Βήμα 4ο Υπολογισμοί

$$Z^* = \frac{\hat{p}-P}{s_{\hat{p}}} = \frac{0,32-0,20}{\sqrt{0,20\cdot 0,80/49}} = 2,10 \qquad z_\alpha = z_{0,10} = 1,282$$

$$p-value = P(\hat{p} > 0,32) = P(\frac{\hat{p}-P}{\sqrt{P\cdot Q/n}} > \frac{0,32-0,20}{\sqrt{0,20\cdot 0,80/49}}) = P(Z^* > 2,10) = 0,0179$$

Βήμα 5ο Στατιστικό συμπέρασμα-απόφαση

Με βάση τον κανόνα απόφασης «Απορρίπτεται η H_0: P≤0,20 αν $Z^* > z_\alpha$» πράγματι απορρίπτουμε την H_0 αφού $Z^*(=2,10) < z_{0,10}(=1,282)$. Με άλλα λόγια είναι πολύ πιθανό η κατάσταση σε όλους τους ΟΤΑ για το υπόψη θέμα πράγματι να έχει χειροτερέψει. Εντούτοις, πάντα διατηρούμε επιφύλαξη, να έχουμε κάνει κάποιο δειγματοληπτικό λάθος.

Εναλλακτικά, με τον κανόνα απόφασης «απορρίπτεται η H_0: P≤0,20 αν (p-value)<α=0,10» επίσης καταλήγουμε στο ίδιο συμπέρασμα να απορρίψουμε τη βασική αφού p-value (=1,79%) < α (=10%).

9. Εκπαιδευτική Ενότητα

• Ανάλυση Διακύμανσης κατά Παράγοντα

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατανοεί και να επεξηγεί τη μέθοδο της Ανάλυσης Διακύμανσης κατά ένα παράγοντα.
- Να εφαρμόζει τη μέθοδο σε συγκεκριμένα δεδομένα.
- Να κατανοεί και να ερμηνεύει τα αποτελέσματα της μεθόδου

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Ανάλυση Διακύμανσης
- Κριτήριο LSD
- Κριτήριο Bonferroni
- Κριτήριο Tukey HSD
- Κριτήριο Scheffe

9.1 Εισαγωγή

Όλα τα μέχρι τώρα στατιστικά τεστ που εξετάσαμε, χρησιμοποιούν ένα ή δύο δείγματα για να επαληθεύσουν υποθέσεις που αφορούν μια ή δυο τιμές της εξαρτημένης (ελεγχόμενης) μεταβλητής. Αν και τα τεστ αυτά χρησιμοποιούνται ευρέως από τους αναλυτές, εντούτοις υπάρχουν και περιπτώσεις όπου καλούμαστε να εξετάσουμε την επίδραση δύο ή περισσοτέρων ελεγχόμενων μεταβλητών με δύο ή και περισσότερες τιμές στην παρατηρούμενη μεταβλητή. Προβλήματα τέτοιου είδους μπορούν να αντιμετωπιστούν με επιτυχία, εφαρμόζοντας τη μέθοδο ανάλυσης της διακύμανσης (Analysis Of Variance –ANOVA) και να ελέγξουμε τη στατιστική σημαντικότητα της διαφοράς των μέσων όταν διαθέτουμε περισσότερα από δύο δείγματα (ομάδες δεδομένων). Ουσιαστικά, η ανάλυση διακύμανσης αποτελεί προέκταση του στατιστικού t-student και μπορεί να εφαρμοστεί είτε διαθέτουμε ανεξάρτητα, είτε εξαρτημένα δείγματα.

Στο κεφάλαιο αυτό, θα αναφερθούμε στην ανάλυση διακύμανσης ενός παράγοντα/κριτηρίου (One-Way Analysis Of Variance), η οποία μας επιτρέπει να συγκρίνουμε μέσους όρους από περισσότερα από δύο δείγματα (ομάδες μιας ανεξάρτητης μεταβλητής). Αντιθέτως, στην περίπτωση που ο έλεγχος αφορά δείγματα που προέρχονται από δύο ανεξάρτητες μεταβλητές, τότε χρησιμοποιούμε την ανάλυση διακύμανσης δύο παραγόντων (Two-Way Analysis Of Variance) κ.ο.κ. Κατά την ανάλυση διακύμανσης ενός παράγοντα/κριτηρίου, η μηδενική υπόθεση δηλώνει ότι τα διαθέσιμα δείγματα/ομάδες προέρχονται από πληθυσμούς που έχουν τον ίδιο μέσο. Αν όμως κατά τη στατιστική ανάλυση απορρίψουμε τη μηδενική υπόθεση και δεχθούμε την εναλλακτική, δηλαδή ότι τα δείγματα διαφέρουν, τότε δεν μπορούμε να ισχυριστούμε ποια συγκεκριμένα δείγματα διαφέρουν μεταξύ τους. Το πρόβλημα αυτό, μπορεί να λυθεί κάπως απλά χρησιμοποιώντας επαναλαμβανόμενους χρονοβόρους και πολλές φορές μη αξιόπιστους ελέγχους με τη στατιστική t-student. Το κενό αυτό, συμπληρώνεται με τη χρησιμοποίηση ειδικών τεστ (τα λεγόμενα post hoc tests) της ανάλυσης διακύμανσης τα οποία μας δίνουν ιδιαίτερη πληροφόρηση σχετικά με τις διαφορές των μέσων μεταξύ των δειγμάτων (αλληλεπιδράσεις).

Για να χρησιμοποιηθεί η ανάλυση διακύμανσης ενός παράγοντα, πρέπει να ισχύουν οι ακόλουθες προϋποθέσεις:

1. Οι διαθέσιμες παρατηρήσεις των δειγμάτων πρέπει να είναι ανεξάρτητες.
2. Οι πληθυσμοί από τους οποίους επιλέξαμε τα δείγματα, θα πρέπει να έχουν την ίδια διακύμανση.
3. Η κατανομή των διαθέσιμων παρατηρήσεων πρέπει να είναι κανονική.

9.2 Ανάλυση διακύμανσης με ένα κριτήριο

9.2.1 Το υπόδειγμα

Όταν στο πλαίσιο μιας πειραματικής έρευνας διαθέτουμε k ομάδες με n_j παρατηρήσεις ανά ομάδα (j=1,..,k), τότε τα δεδομένα του δείγματος μπορούν να ταξινομηθούν σε ένα πίνακα με την ακόλουθη μορφή:

Πίνακας 9.1 Διάταξη παρατηρήσεων

	Ομάδες (Μεταχειρίσεις)					
	1	**2**	**3**		k	
X_{ij}	X_{11}	X_{12}	X_{13}	...	X_{1k}	
	X_{21}	X_{22}	X_{23}	...	X_{2k}	
	X_{31}	X_{32}	X_{33}	...	X_{3k}	
	.	.	.	...	.	
	.	.	.	...	.	
	.	.	.	...	.	
	$X_{n_1 1}$	$X_{n_2 2}$	$X_{n_3 3}$	...	$X_{n_k k}$	
Μεγέθη δειγμάτων	n_1	n_2	n_3		n_k	$N=\sum_{j=1}^{k} n_j$
Σύνολα	$T_{\cdot 1}$	$T_{\cdot 2}$	$T_{\cdot 3}$	...	$T_{\cdot k}$	$T_{\cdot\cdot}=\sum_{j=1}^{k} T_{\cdot j}$
Μέσοι όροι	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	$\bar{X}_{\cdot 3}$		$\bar{X}_{\cdot k}$	$\bar{X}_{\cdot\cdot}=\frac{T_{\cdot\cdot}}{N}$

Όπου:

X_{ij} = η i η παρατήρηση της j ομάδας για i-1,....,n_j και j-1,...,k.

$T_{\cdot j}=\sum_{i=1}^{n} X_{ij}$ = άθροισμα των παρατηρήσεων της j στήλης.

$\bar{X}_{\cdot j}=\frac{T_{\cdot j}}{n_j}$ = αριθμητικός μέσος της j στήλης.

$T_{\cdot\cdot}=\sum_{j=1}^{k} T_{\cdot j}=\sum_{j=1}^{k}\sum_{i=1}^{n} X_{ij}$ = άθροισμα όλων των παρατηρήσεων.

$\bar{X}_{..} = \frac{T_{..}}{N}$ = γενικός αριθμητικός μέσος όλων των παρατηρήσεων.

Οι υποθέσεις που διέπουν το υπόδειγμα είναι:

1. Οι παρατηρήσεις X_{ij} κάθε ομάδας αποτελούν k ανεξάρτητα δείγματα από αντίστοιχους πληθυσμούς.

2. Καθένας από τους k πληθυσμούς ακολουθεί την κανονική κατανομή με μέση τιμή μ_j και κοινή διακύμανση σ^2, για j=1,....,k.

3. Οι επιδράσεις των ομάδων (μεταχειρίσεων) τ_j είναι σταθεροί αριθμοί που ικανοποιούν τη σχέση $\sum_{j=1}^{k} \tau_j = 0$.

Από το σημείο αυτό, ας ξεκινήσουμε με ένα παράδειγμα το οποίο θα χρησιμοποιήσουμε και στην υπόλοιπη ανάλυση διακύμανσης.

Παράδειγμα 9.1

Μια επιχείρηση που δραστηριοποιείται στον κλάδο τροφίμων και ποτών, μεταξύ των άλλων προϊόντων, παράγει και φυσικούς χυμούς φρούτων σε συσκευασία του 1 λίτρου. Ο διευθυντής πωλήσεων της επιχείρησης, υλοποιώντας ένα σχέδιο έρευνας αγοράς για την προώθηση των πωλήσεων των φυσικών χυμών, ενδιαφέρεται να εξετάσει την κατανάλωση φυσικών χυμών (σε λίτρα) μιας τετραμελούς οικογένειας κατά τη διάρκεια ενός μηνός. Για το σκοπό αυτό, συλλέγει στοιχεία από τυχαίο δείγμα 25 οικογενειών που κατοικούν σε διαφορετικές πόλεις (Α, Β, Γ, Δ) και με ισόποση κατανομή σ' αυτές (δηλαδή 5 στην πόλη Α, 5 στη πόλη Β κ.λπ.). Τα στοιχεία τα οποία συνέλεξε μετά το πέρας της έρευνας παρουσιάζονται στον ακόλουθο πίνακα:

Οικογένεια	Πόλη Α	Πόλη Β	Πόλη Γ	Πόλη Δ
1	18,5	26,3	20,6	25,4
2	24,0	25,3	25,2	19,9
3	17,2	24,0	20,8	22,6
4	19,9	21,2	24,7	17,5
5	18,0	24,5	22,9	20,4

Να κατασκευαστεί ο πίνακας διάταξης των παρατηρήσεων του πλήρως τυχαιοποιημένου σχεδίου και να υπολογιστεί η γενική μέση τιμή των παρατηρήσεων (κατανάλωσης φυσικού χυμού).

Απάντηση:

Ο πίνακας διάταξης των παρατηρήσεων του πλήρως τυχαιοποιημένου σχεδίου έχει ως εξής:

Πίνακας 9.2 Κατανάλωση φυσικού χυμού (σε λίτρα)

	Πόλεις				
Οικογένεια	**Πόλη Α**	**Πόλη Β**	**Πόλη Γ**	**Πόλη Δ**	
1	X_{11}=18,5	X_{12}=26,3	X_{13}=20,6	X_{14}=25,4	
2	X_{21}=24,0	X_{22}=25,3	X_{23}=25,2	X_{24}=19,9	
3	X_{31}=17,2	X_{32}=24,0	X_{33}=20,8	X_{34}=22,6	
4	X_{41}=19,9	X_{42}=21,2	X_{43}=24,7	X_{44}=17,5	
5	X_{51}=18,0	X_{52}=24,5	X_{53}=22,9	X_{55}=20,4	
Μεγέθη δειγμάτων	n_1=5	n_2=5	n_3=5	n_4=5	$N=\sum_{i=1}^{4} n_i$ =20
Σύνολα	$T_{.1}$=97,6	$T_{.2}$=121,3	$T_{.3}$=114,2	$T_{.4}$=105,8	$T_{..}=\sum_{i=1}^{4} T_{.i}$ =109,73
Μέσοι όροι	$\bar{X}_{.1}$=19,52	$\bar{X}_{.2}$=24,26	$\bar{X}_{.3}$=22,84	$\bar{X}_{.4}$=21,16	$\bar{X}_{..}=\frac{T_{..}}{N}$=21,945

Στο παράδειγμα μας, έχουμε k=4 ομάδες (πόλεις) με πλήθος παρατηρήσεων n=5 σε κάθε ομάδα (πόλη) έτσι ώστε το συνολικό πλήθος των παρατηρήσεων να είναι N=k·n=4·5=20.

$$\bar{X}_{..}=\frac{T_{..}}{N}=\frac{T_{.1}+T_{.2}+T_{.3}+T_{.4}}{20}=\frac{97,6+121,3+114,2+105,8}{20}=\frac{109,73}{20}=21,945$$

ή από τη σχέση

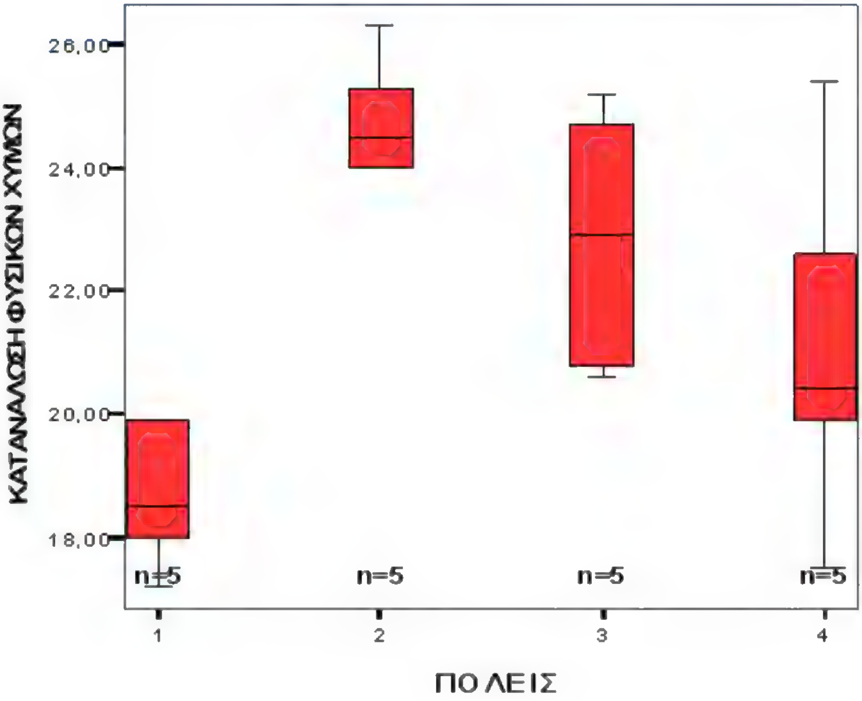
$$\bar{X}=\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}}{n}=\frac{18,5+24,0+\ldots+26,3+25,3+\ldots+20,6+\ldots+25,4+\ldots+20,4}{20}=21,945$$

Ο γενικός μέσος όρος που υπολογίστηκε, σημαίνει ότι η μέση κατανάλωση φυσικού χυμού του συνόλου των οικογενειών που ρωτήθηκαν, ανέρχεται στα 21,945 λίτρα μηνιαίως.

9.2.2 Έλεγχος υποθέσεων

Η όλη παραπάνω διεργασία της διάταξης των παρατηρήσεων σε πίνακα, μπορούμε να πούμε ότι αποτελεί το πρώτο στάδιο για την ανάλυση διακύμανσης ενός προβλήματος. Επιπλέον, η πινακοποίηση των δεδομένων μας με την παραπάνω μορφή, μας βοηθά στην κατανόηση του προβλήματος αλλά και στην παροχή κάποιων πρώτων βασικών περιγραφικών μέτρων (γενικός και ανά ομάδα (πόλη) μέσος όρος κατανάλωσης φυσικού χυμού).

Επειδή, απώτερος στόχος της ανάλυσης διακύμανσης είναι ο καθορισμός του στατιστικού κριτηρίου για τον έλεγχο της μηδενικής υπόθεσης, για να φτάσουμε εκεί θα πρέπει ν'ασχοληθούμε με το θέμα της μεταβλητότητας των παρατηρήσεων (μέσα σε κάθε δείγμα, μεταξύ των δειγμάτων και συνολικά μεταξύ των παρατηρήσεων του συνολικού δείγματος). Ο όρος μεταβλητότητα αναφέρεται στο άθροισμα των τετραγώνων των αποκλίσεων των παρατηρήσεων από την μέση τιμή τους και ονομάζεται άθροισμα τετραγώνων (Sum of Squares-SS). Παρατηρώντας το επόμενο διάγραμμα, βλέπουμε ότι η μεταβλητικότητα της κατανάλωσης χυμού είναι εμφανής, τόσο μέσα σε κάθε πόλη, όσο και μεταξύ των πόλεων.



Προς την κατεύθυνση αυτή, θα συνεχίσουμε με τη διατύπωση των υποθέσεων του υποδείγματος. Σ' ένα υπόδειγμα ανάλυσης διακύμανσης και ειδικότερα στη γενική του μορφή, ο έλεγχος υποθέσεων διατυπώνεται ως εξής:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (Ολοι οι μέσοι είναι ίσοι μεταξύ τους) $H_r : \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$ (Ολοι οι μέσοι δεν είναι ίσοι μεταξύ τους)	**(9.1)**

Για το παράδειγμα μας, οι υποθέσεις διατυπώνονται ανάλογα ως ακολούθως:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (Η μέση κατανάλωση φυσικού χυμού είναι ίδια σε όλες τις πόλεις)

$H_r : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ (Η μέση κατανάλωση φυσικού χυμού διαφέρει από πόλη σε πόλη)

Από την παράθεση των στοιχείων της έρευνας (Πίνακας 9.2), το πρόβλημα της μεταβλητότητας (διασποράς) των παρατηρήσεων είναι εμφανές. Η μεταβλητότητα αυτή εντοπίζεται:

α) Μέσα σε κάθε δείγμα. Για παράδειγμα, οι οικογένειες στην πόλη Α, καταναλώνουν μεταξύ τους διαφορετική ποσότητα φυσικών χυμών. Το ίδιο φαινόμενο παρατηρείται και στις υπόλοιπες πόλεις.

β) Μεταξύ των δειγμάτων. Έντονη είναι και η διαφορά μεταξύ της μέσης κατανάλωσης φυσικών χυμών από πόλη σε πόλη.

γ) Στο σύνολο των παρατηρήσεων του γενικού δείγματος της έρευνας. Αυτό προκύπτει εύκολα, αν λάβουμε υπόψη μας τον γενικό μέσο που υπολογίσαμε και τον συγκρίνουμε (ή υπολογίσουμε τις αποκλίσεις) με τις παρατηρήσεις των δειγμάτων.

Εξάλλου, η ανάλυση διακύμανσης ορίζεται σαν μια διαδικασία κατά την οποία η συνολική μεταβλητότητα που υπάρχει στα δεδομένα, διασπάται σε επιμέρους συνιστώσες που οφείλονται σε διαφορετικές πηγές προέλευσης. Επομένως, εύκολα μπορούμε να ορίσουμε τις έννοιες αυτές:

- $SS_{total} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_m)^2$ = Ολικό άθροισμα τετραγώνων (Sum of Squares Total) των αποκλίσεων των παρατηρήσεων από τον γενικό μέσο ή συνολική διασπορά των δειγματικών τιμών. Όπου, με το $\sum_{i=1}^{n_j}$ αθροίζουμε τις τετραγωνισμένες αποκλίσεις μέσα σε κάθε ομάδα, ενώ με το $\sum_{j=1}^{k}$ αθροίζουμε τα αποτελέσματα των ομάδων.

- $SS_{\text{within}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{0j})^2$ = Άθροισμα τετραγώνων των αποκλίσεων των

παρατηρήσεων κάθε δείγματος από τον μέσο τους ή άθροισμα τετραγώνων μέσα στα δείγματα/ομάδες (within groups Sum of Squares).

- $SS_{\text{between}} = \sum_{j=1}^{k} n_j(\bar{X}_{0j} - \bar{X}_{00})^2$ = Άθροισμα Τετραγώνων των αποκλίσεων των μέσων

των δειγμάτων από τον γενικό μέσο ή άθροισμα μεταξύ των δειγμάτων/ομάδων (between groups sum of squares).

Επομένως, ισχύει:

$SS_{\text{total}} = SS_{\text{within}} + SS_{\text{between}}$	**(9.2)**

Εν συνεχεία, αν διαιρέσουμε τα αθροίσματα τετραγώνων των συνιστωσών της συνολικής μεταβλητότητας με τους αντίστοιχους βαθμούς ελευθερίας τους, τότε προκύπτει το μέσο άθροισμα τετραγώνων των αποκλίσεων. Δηλαδή:

- $MSS_{\text{between}} = \dfrac{SS_{\text{between}}}{k-1}$ = Μέσο άθροισμα τετραγωνικών αποκλίσεων μεταξύ των

δειγμάτων (Mean of the sum of squared deviations between groups).

- $MSS_{\text{within}} = \dfrac{SS_{\text{within}}}{n-k}$ = Μέσο άθροισμα τετραγωνικών αποκλίσεων εντός των δειγμάτων

(Mean of the sum of squared deviations within groups).

Τέλος, το κριτήριο ελέγχου της ισότητας των μέσων, είναι η στατιστική F και προκύπτει από τον λόγο:

$F_{(k-1),(n-k)} = \dfrac{MSS_{\text{between}}}{MSS_{\text{within}}} = \dfrac{\frac{SS_{\text{between}}}{k-1}}{\frac{SS_{\text{within}}}{n-k}}$	**(9.3)**

Επομένως, αν για επίπεδο σημαντικότητας α $F_{(k-1),(n-k)} < F_{(k-1),(n-k),a}$, τότε δεχόμαστε την μηδενική υπόθεση ότι οι μέσοι των δειγμάτων/ομάδων είναι ίσοι μεταξύ τους, διαφορετικά δεχόμαστε την εναλλακτική υπόθεση.

Τέλος, κατασκευάζουμε τον πίνακα ανάλυσης διακύμανσης στον οποίο απεικονίζονται συνοπτικά, όλοι οι απαιτούμενοι υπολογισμοί για τον προσδιορισμό του στατιστικού κριτηρίου F.

Πίνακας 9.3 Ανάλυσης διακύμανσης

Πηγή μεταβλητότητας (source of variation)	Άθροισμα τετραγώνων αποκλίσεων (Sum of Squares)	Βαθμοί ελευθερίας (Degrees of Freedom) d.f	Μέσο τετραγωνικό σφάλμα (mean square)	Στατιστική F
Μεταξύ ομάδων	$SS = \sum n(\bar{X} - \bar{X})$	$k\text{-}1$	$MSS_{between} = \frac{SS_{between}}{k-1}$	
Εντός των ομάδων	$SS_{within} = \sum\sum(X - \bar{X})^2$	$N\text{-}k$	$MSS_{within} = \frac{SS_{within}}{n-k}$	$F = \frac{MSS}{MSS}$
Συνολική (Total)	$SS_{total} = \sum\sum(X - \bar{X})^2$	$N\text{-}1$		

Παράδειγμα 9.2

Με τα δεδομένα του παραδείγματος 9.1 και τα στοιχεία του πίνακα 1, να ελεγχθεί σε επίπεδο σημαντικότητας α=0,05 η μηδενική υπόθεση, δηλαδή ότι η μέση κατανάλωση φυσικού χυμού είναι ίδια σε όλες τις πόλεις.

Απάντηση:

Ο έλεγχος θα γίνει με το στατιστικό κριτήριο F. Για να μπορέσουμε να εκτιμήσουμε την τιμή του, θα πρέπει κατ' αρχάς να υπολογίσουμε το άθροισμα τετραγώνων των αποκλίσεων, εντός, μεταξύ και στο σύνολο των πόλεων, δηλαδή να βρούμε τις πηγές της συνολικής μεταβλητικότητας/διασποράς των παρατηρήσεων του συνολικού δείγματος. Επομένως:

Άθροισμα τετραγώνων των αποκλίσεων εντός των πόλεων:

$$SS_{within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\eta j})^2$$

$$=(18{,}5-19{,}52)^2+\ldots+(18-19{,}52)^2+(26{,}3-24{,}26)^2+\ldots+(24{,}5-24{,}26)^2$$

$$+(20{,}6-22{,}84)^2+\ldots+(22{,}9-22{,}84)^2+(25{,}4-21{,}16)^2+\ldots+(20{,}4-21{,}16)^2$$

$$=97{,}540$$

Άθροισμα τετραγώνων των αποκλίσεων μεταξύ των πόλεων:

$$SS_{\text{between}} = \sum_{j=1}^{k} n_j (\overline{X}_{\text{m}j} - \overline{X}_{\text{m}})^2$$

$$=5 \cdot (19{,}52\text{-}21{,}945)^2 + 5 \cdot (24{,}26\text{-}21{,}945)^2 + 5 \cdot (22{,}84\text{-}21{,}945)^2$$

$$=+5 \cdot (21{,}16\text{-}21{,}945)^2 = 63{,}286$$

Ολικό Άθροισμα τετραγώνων των αποκλίσεων:

$$SS_{\text{total}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{\text{m}})^2$$

$$= (18{,}5\text{-}21{,}945)^2 + (24\text{-}21{,}945)^2 + \ldots + (20{,}4\text{-}21{,}945)^2 = 160{,}790$$

Εν συνεχεία, υπολογίζουμε το μέσο άθροισμα τετραγώνων των αποκλίσεων εντός, μεταξύ και στο σύνολο των πόλεων. Επομένως:

Μέσο άθροισμα τετραγώνων των αποκλίσεων εντός των πόλεων:

$$MSS_{\text{within}} = \frac{SS_{\text{within}}}{n-k} = \frac{97{,}504}{20-4} = 6{,}094$$

Το αποτέλεσμα αυτό, έχει ιδιαίτερο ενδιαφέρον και σημαίνει ότι κατά μέσο όρο, η μέση οικογενειακή κατανάλωση φυσικού χυμού στις πόλεις Α, Β, Γ και Δ, διαφέρει/αποκλίνει από την αντίστοιχη μέση οικογενειακή κατανάλωση που παρατηρήθηκε στις πόλεις αυτές κατά 6,1 λίτρα φυσικού χυμού περίπου.

Μέσο άθροισμα τετραγωνικών αποκλίσεων μεταξύ των πόλεων:

$$MSS_{\text{between}} = \frac{SS_{\text{between}}}{k-1} = \frac{63{,}286}{4-1} = 21{,}096$$

Αυτό σημαίνει ότι κατά μέσο όρο, η μέση οικογενειακή κατανάλωση φυσικού χυμού που παρατηρήθηκε στις πόλεις Α, Β, Γ, Δ, διαφέρει/αποκλίνει από τη γενική μέση κατανάλωση του συνόλου των οικογενειών του δείγματος (που ζουν και στις τέσσερις πόλεις) κατά 0,85 λίτρα (21,945-21,096) φυσικού χυμού.

Μετά τους παραπάνω υπολογισμούς και διαιρώντας τις μέσες τετραγωνικές αποκλίσεις μεταξύ των πόλεων με τις αντίστοιχες εντός των πόλεων, προκύπτει η τιμή του στατιστικού κριτηρίου F. Δηλαδή:

$$F_{(k-1),(n-k)} = \frac{MSS_{between}}{MSS_{within}} = \frac{\frac{SS_{between}}{k-1}}{\frac{SS_{within}}{n-k}} = \frac{21,095}{6,094} = 3,462$$

Τέλος, για τη λήψη απόφασης σχετικά με τον διενεργούμενο έλεγχο, ελέγχουμε αν $F_{(k-1),(n-k)} < F_{(k-1),(n-k),a}$ σε επίπεδο σημαντικότητας α=0,05.

Επομένως: $F_{(k-1),(n-k)} = 3,462 > F_{3,16,0.05} = 3,24$ που σημαίνει ότι απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική, δηλαδή συμπεραίνουμε ότι η μέση κατανάλωση φυσικών χυμών τουλάχιστον μιας πόλης διαφέρει από τις αντίστοιχες των υπολοίπων πόλεων του δείγματος.

Το ερώτημα όμως που τίθεται τώρα είναι: «Ποιες είναι αυτές οι πόλεις που η μέση κατανάλωσή τους διαφέρει σε σχέση με τις υπόλοιπες;» Το ερώτημα αυτό, μπορούμε να το απαντήσουμε, αν χρησιμοποιήσουμε κατάλληλα κριτήρια τα οποία μας επιτρέπουν να συγκρίνουμε τη μέση κατανάλωση της κάθε πόλης με όλες τις άλλες. Τα κριτήρια αυτά όπως αναφέραμε ονομάζονται «post hoc κριτήρια πολλαπλών συγκρίσεων» και η χρήση τους αποφασίζεται μετά την ανάλυση διακύμανσης.
Στην επόμενη παράγραφο, θα παρουσιάσουμε σύντομα μερικά από τα κυριότερα κριτήρια πολλαπλών συγκρίσεων που χρησιμοποιούνται ευρέως στην πράξη και τα οποία χρησιμοποιήσαμε στα δεδομένα του προβλήματος για να εντοπίσουμε επιμέρους διαφορές στην κατανάλωση μεταξύ των πόλεων.

9.2.3 Κριτήρια πολλαπλών συγκρίσεων (post hoc tests)

9.2.3.1 Κριτήριο LSD

Το κριτήριο αυτό, ονομάζεται κριτήριο της ελάχιστης σημαντικής διαφοράς ή προστατευμένο t. Χρησιμοποιείται όταν η τιμή της στατιστικής F που υπολογίστηκε κατά την ανάλυση διακύμανσης είναι στατιστικά σημαντική. Ουσιαστικά, πρόκειται για τροποποίηση του στατιστικού t-test και υπολογίζεται από τον τύπο:

$t = \dfrac{\bar{X}_{0j} - \bar{X}_{0j'}}{\sqrt{MSS_{within}\left(\dfrac{1}{n_j} + \dfrac{1}{n_{j'}}\right)}}$ για $j \neq j'$	**(9.4)**

Όπου:

$\overline{X}_{\Omega j}$ και $\overline{X}_{\Omega j'}$ = οι μέσοι όροι δύο ομάδων που συγκρίνονται.

n_j και $n_{j'}$ = το πλήθος των τιμών κάθε ομάδας που συγκρίνεται.

MSS_{within} = το μέσο άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των ομάδων.

Οι βαθμοί ελευθερίας για το κριτήριο LSD βρίσκονται από τον πίνακα ανάλυσης διακύμανσης ($df = df_{\text{within}}$).

Παράδειγμα 9.3

Με τα δεδομένα του παραδείγματος 9.1 και τα στοιχεία του πίνακα 1, να ελεγχθεί σε επίπεδο σημαντικότητας α=0,05 η μηδενική υπόθεση ότι η μέση κατανάλωση φυσικού χυμού μεταξύ των πόλεων Α και Β είναι ίση, δηλαδή έχουν την ίδια μέση κατανάλωση, με το κριτήριο LSD.

Απάντηση:

Με βάση τον τύπο υπολογισμού του κριτηρίου LSD, έχουμε:

$$t = \frac{\overline{X}_{\Omega j} - \overline{X}_{\Omega j'}}{\sqrt{MSS_{\text{within}}\left(\frac{1}{n_j}+\frac{1}{n_{j'}}\right)}} = \frac{\overline{X}_{\Omega 1} - \overline{X}_{\Omega 2}}{\sqrt{MSS_{\text{within}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{19,52-24,26}{\sqrt{6,094\cdot\left(\frac{1}{5}+\frac{1}{5}\right)}} = \frac{-4,74}{1,103993} = -4,2935$$

που σημαίνει ότι η παρατηρούμενη μέση διαφορά κατανάλωσης (-4,74) στις πόλεις Α και Β είναι στατιστικά σημαντική (δηλ. η μέση κατανάλωση είναι διαφορετική μεταξύ των δύο αυτών πόλεων). Όλες οι δυνατές συγκρίσεις μέσης κατανάλωσης φυσικών χυμών μεταξύ των πόλεων, παρουσιάζονται στον πίνακα που ακολουθεί:

Πίνακας 9.4 ΠΟΛΛΑΠΛΕΣ ΣΥΓΚΡΙΣΕΙΣ ΜΕ ΤΟ ΚΡΙΤΗΡΙΟ LSD ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: ΚΑΤΑΝΑΛΩΣΗ ΦΥΣΙΚΟΥ ΧΥΜΟΥ

Κριτήριο	(I) ΠΟΛΕΙΣ	(J) ΠΟΛΕΙΣ	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	-4,74*	1,561282	0,007866	-8,050	-1,430
		3	-3,32*	1,561282	0,049376	-6,630	-0,010
		4	-1,64	1,561282	0,309133	-4,950	1,670
	2	1	4,74*	1,561282	0,007866	1,430	8,050
		3	1,42	1,561282	0,376579	-1,890	4,730
		4	3,1	1,561282	0,064497	-0,210	6,410
	3	1	3,32*	1,561282	0,049376	0,010	6,630
		2	-1,42	1,561282	0,376579	-4,730	1,890
		4	1,68	1,561282	0,297868	-1,630	4,990
	4	1	1,64	1,561282	0,309133	-1,670	4,950
		2	-3,1	1,561282	0,064497	-6,410	0,210
		3	-1,68	1,561282	0,297868	-4,990	1,630

* The mean difference is significant at the .05 level.

9.2.3.2 Κριτήριο Bonferroni

Το κριτήριο αυτό, μπορεί να χρησιμοποιηθεί ακόμα και όταν η τιμή της στατιστικής F που υπολογίστηκε κατά την ανάλυση διακύμανσης δεν είναι στατιστικά σημαντική. Οι συγκρίσεις μεταξύ των ομάδων, γίνονται με τη χρήση του κριτηρίου LSD. Η μόνη διαφορά που υπάρχει μεταξύ τους είναι, ότι η σημαντικότητα της τιμής του κριτηρίου t που υπολογίζεται, θα πρέπει να ελεγχθεί σε επίπεδο σημαντικότητας α/πλήθος συγκρίσεων. Στη περίπτωσή μας, το επίπεδο σημαντικότητας του α θα είναι α=0,05/12=0,0042.

Όλες οι δυνατές συγκρίσεις μέσης κατανάλωσης φυσικών χυμών μεταξύ των πόλεων, παρουσιάζονται στον πίνακα που ακολουθεί:

Πίνακας 9.5 ΠΟΛΛΑΠΛΕΣ ΣΥΓΚΡΙΣΕΙΣ ΜΕ ΤΟ ΚΡΙΤΗΡΙΟ Bonferroni ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: ΚΑΤΑΝΑΛΩΣΗ ΦΥΣΙΚΟΥ ΧΥΜΟΥ

	(I) ΠΟΛΕΙΣ	(J) ΠΟΛΕΙΣ	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
Κριτήριο						Lower Bound	Upper Bound
Bonferroni	1	2	-4,74*	1,561282	0,047197	-9,437	-0,043
		3	-3,32	1,561282	0,296255	-8,017	1,377
		4	-1,64	1,561282	1	-6,337	3,057
	2	1	4,74*	1,561282	0,047197	0,043	9,437
		3	1,42	1,561282	1	-3,277	6,117
		4	3,1	1,561282	0,386979	-1,597	7,797
	3	1	3,32	1,561282	0,296255	-1,377	8,017
		2	-1,42	1,561282	1	-6,117	3,277
		4	1,68	1,561282	1	-3,017	6,377
	4	1	1,64	1,561282	1	-3,057	6,337
		2	-3,1	1,561282	0,386979	-7,797	1,597
		3	-1,68	1,561282	1	-6,377	3,017

* The mean difference is significant at the .05 level.

9.2.3.3 Κριτήριο Tukey HSD

Ονομάζεται και ως κριτήριο «ειλικρινούς σημαντικής διαφοράς». Θεωρείται από τα πλέον ασφαλή κριτήρια πολλαπλών συγκρίσεων. Πρόκειται για πολύ συντηρητικό κριτήριο, με αποτέλεσμα κάποιες ελεγχόμενες διαφορές να μην εμφανίζονται ως στατιστικά σημαντικές, ενώ αν εφαρμόζαμε κάποιο άλλο κριτήριο πιθανόν οι διαφορές αυτές να προέκυπταν σημαντικές. Κατά τη διαδικασία του ελέγχου αυτού, θα πρέπει να προσέξουμε ότι όταν έχουμε να συγκρίνουμε δύο μέσους, τότε αφαιρούμε πάντα τον μικρότερο από τον μεγαλύτερο. Ο δείκτης q του κριτηρίου, υπολογίζεται από τον τύπο:

$$q = \frac{\bar{X}_{q j} - \bar{X}_{q j'}}{\sqrt{\frac{MSS_{within}}{n}}} \quad \text{για } j \neq j' \qquad (9.5)$$

Όπου:

$\bar{X}_{ij}$ = ο μεγαλύτερος από τους δύο μέσους όρους που συγκρίνονται.

$\bar{X}_{ij'}$ = ο μικρότερος από τους δύο μέσους όρους που συγκρίνονται.

n= το πλήθος των τιμών κάθε ομάδας (ομάδες με ίδιο αριθμό παρατηρήσεων)

MSS_{within} = το μέσο άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των ομάδων.

Ο παραπάνω δείκτης, εφαρμόζεται μόνο όταν οι ομάδες έχουν τον ίδιο αριθμό παρατηρήσεων. Διαφορετικά, όταν το πλήθος είναι άνισο, τότε η τιμή του αντικαθίσταται από τον αρμονικό μέσο όρο $\tilde{n} = \dfrac{k}{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}+\ldots+\dfrac{1}{n_k}\right)}$.

Ο έλεγχος της μηδενικής υπόθεσης στην περίπτωση αυτή, γίνεται αφού πρώτα βρούμε τους βαθμούς ελευθερίας (στη περίπτωσή μας $df = df_{within} = 16$) και εν συνεχεία συγκρίνουμε την τιμή που υπολογίσαμε με την κρίσιμη τιμή. Όλες οι δυνατές συγκρίσεις μέσης κατανάλωσης φυσικών χυμών μεταξύ των πόλεων, παρουσιάζονται στον πίνακα που ακολουθεί:

Πίνακας 9.6 ΠΟΛΛΑΠΛΕΣ ΣΥΓΚΡΙΣΕΙΣ ΜΕ ΤΟ ΚΡΙΤΗΡΙΟ Tukey HSD ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: ΚΑΤΑΝΑΛΩΣΗ ΦΥΣΙΚΟΥ ΧΥΜΟΥ

Κριτήριο	(I) ΠΟΛΕΙΣ	(J) ΠΟΛΕΙΣ	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1	2	-4,74*	1,561282	0,035628	-9,207	-0,273
		3	-3,32	1,561282	0,186871	-7,787	1,147
		4	-1,64	1,561282	0,723193	-6,107	2,827
	2	1	4,74*	1,561282	0,035628	0,273	9,207
		3	1,42	1,561282	0,800098	-3,047	5,887
		4	3,1	1,561282	0,234071	-1,367	7,567
	3	1	3,32	1,561282	0,186871	-1,147	7,787
		2	-1,42	1,561282	0,800098	-5,887	3,047
		4	1,68	1,561282	0,708491	-2,787	6,147
	4	1	1,64	1,561282	0,723193	-2,827	6,107
		2	-3,1	1,561282	0,234071	-7,567	1,367
		3	-1,68	1,561282	0,708491	-6,147	2,787

* The mean difference is significant at the .05 level.

9.2.3.4 Κριτήριο Scheffe

Πρόκειται επίσης για συντηρητικό κριτήριο το οποίο βασίζεται στη στατιστική F και υπολογίζεται από τη σχέση:

$$t = \frac{\bar{X}_{\eta j} - \bar{X}_{\eta j'}}{\sqrt{MSS_{within}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}} \quad \text{για } j \neq j' \qquad (9.6)$$

Όπου:

$\bar{X}_{\eta j}$ και $\bar{X}_{\eta j'}$ = οι μέσοι όροι δύο ομάδων που συγκρίνονται.

n_j και n_j = το πλήθος των τιμών κάθε ομάδας που συγκρίνεται.

MSS_{within} = το μέσο άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των ομάδων.

Όλες οι δυνατές συγκρίσεις μέσης κατανάλωσης φυσικών χυμών μεταξύ των πόλεων, παρουσιάζονται στον πίνακα που ακολουθεί:

Πίνακας 9.7 ΠΟΛΛΑΠΛΕΣ ΣΥΓΚΡΙΣΕΙΣ ΜΕ ΤΟ ΚΡΙΤΗΡΙΟ Scheffe, ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ: ΚΑΤΑΝΑΛΩΣΗ ΦΥΣΙΚΟΥ ΧΥΜΟΥ

Κριτήριο	(I) ΠΟΛΕΙΣ	(J) ΠΟΛΕΙΣ	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Scheffe	1	2	-4,74	1,561282	0,057762	-9,607	0,127
		3	-3,32	1,561282	0,250748	-8,187	1,547
		4	-1,64	1,561282	0,777239	-6,507	3,227
	2	1	4,74	1,561282	0,057762	-0,127	9,607
		3	1,42	1,561282	0,84204	-3,447	6,287
		4	3,1	1,561282	0,304352	-1,767	7,967
	3	1	3,32	1,561282	0,250748	-1,547	8,187
		2	-1,42	1,561282	0,84204	-6,287	3,447
		4	1,68	1,561282	0,764603	-3,187	6,547
	4	1	1,64	1,561282	0,777239	-3,227	6,507
		2	-3,1	1,561282	0,304352	-7,967	1,767
		3	-1,68	1,561282	0,764603	-6,547	3,187

10. Εκπαιδευτική Ενότητα

- **Ανάλυση Παλινδρόμησης και Συσχέτισης**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατανοεί την έννοια της συσχέτισης.
- Να κατανοεί και να ερμηνεύει τα αποτελέσματα της μεθόδου σε απλά και σύνθετα δεδομένα.
- Να εφαρμόζει απλή και πολλαπλή παλινδρόμηση.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Εξαρτημένη μεταβλητή (Y)
- Ανεξάρτητη μεταβλητή (X)
- Συντελεστής συσχέτισης
- Γραμμική συσχέτιση (linear correlation)
- Μη γραμμική συσχέτιση (non linear correlation)
- Backward Elimination
- Forward Procedure
- Stepwise Regression

10.1 Εισαγωγή

Ο κυριότερος σκοπός πολλών στατιστικών αναλύσεων, συνίσταται στη διερεύνηση της σχέσης μεταξύ δύο ή και περισσοτέρων μεταβλητών που αντιπροσωπεύουν ποσοτικά χαρακτηριστικά μονάδων ενός πληθυσμού (ή ενός δείγματος εκ του πληθυσμού). Για παράδειγμα, έστω ότι θέλουμε να μάθουμε κάτι για τη σχέση μεταξύ, οικογενειακού εισοδήματος και επιπέδου μόρφωσης των κατοίκων μιας περιοχής ή/και των δαπανών τους για θέματα εκπαίδευσης, μηνιαίων αποδοχών και ηλικίας ή/και έτη εμπειρίας ή/και επίπεδο εκπαίδευσης εργαζομένων στον ξενοδοχειακό κλάδο, δαπάνης για αγορά ενός καλλυντικού (π.χ. κρέμας νυκτός) και οικογενειακού εισοδήματος ή/και της διαφημιστικής δαπάνης των επιχειρήσεων παραγωγής του, των πωλήσεων μιας αντιπροσωπίας αυτοκινήτων και του διαθέσιμου αριθμού πωλητών της ή/και της διαφημιστικής δαπάνης, της ζήτησης ενός προϊόντος και της μέσης τιμής διάθεσής του στην αγορά από τον ανταγωνισμό κ.ά. Θεωρώντας τα παραπάνω ποσοτικά χαρακτηριστικά ως μεταβλητές ενός προβλήματος, μπορούμε να πούμε ότι η διερεύνηση της φύσης και της έντασης στη σχέση δύο ή περισσοτέρων μεταβλητών, μπορεί να γίνει με τη μέθοδο της παλινδρόμησης και της συσχέτισης.

Ειδικότερα, με την παλινδρόμηση εκτιμούμε τη σχέση μιας μεταβλητής (εξαρτημένης ή απόκρισης) ως προς μία άλλη ή άλλες (ανεξάρτητες, ή ελεγχόμενες ή επεξηγηματικές), εκφράζοντας τη μία μεταβλητή ως γραμμική συνάρτηση ή όχι της άλλης ή των άλλων μεταβλητών. Κυριότερος σκοπός της μεθόδου της παλινδρόμησης είναι η εκτίμηση και η πρόβλεψη των τιμών της μίας μεταβλητής μέσω των τιμών της άλλης (ή των άλλων). Αντίθετα, στη περίπτωση που αδυνατούμε να θεωρήσουμε τη μία μεταβλητή ως εξαρτημένη και την άλλη (ή τις άλλες) ως ανεξάρτητη, εφαρμόζουμε τη μέθοδο της συσχέτισης με βάση την οποία μετράμε τον βαθμό συμμεταβολής δύο ή περισσοτέρων μεταβλητών ή την ένταση της γραμμικής σχέσης που πιθανόν να υπάρχει μεταξύ τους.
Στην ενότητα αυτή θα παρουσιάσουμε τις δύο αυτές μεθόδους, διερευνώντας τη φύση και την ένταση των χαρακτηριστικών ενός προβλήματος με δύο και περισσότερες μεταβλητές. Επομένως, στη πρώτη περίπτωση θα λέμε ότι αναφερόμαστε στην απλή παλινδρόμηση και απλή συσχέτιση, ενώ στη δεύτερη στην πολλαπλή παλινδρόμηση και πολλαπλή συσχέτιση.

10.2 Απλή γραμμική παλινδρόμηση

Όπως αναφέρθηκε προηγουμένως, η απλή παλινδρόμηση εφαρμόζεται όταν εξετάζουμε τη φύση και την ένταση δύο χαρακτηριστικών (μεταβλητών) Χ και Ψ ενός προβλήματος. Η μεταβλητή Χ ονομάζεται συνήθως ανεξάρτητη μεταβλητή και είναι αυτή την οποία μπορούμε να ελέγξουμε, δηλαδή οι τιμές της επιλέγονται/καθορίζονται από τον ερευνητή (π.χ. το ύψος της διαφημιστικής δαπάνης ενός προϊόντος, το μέγεθος κατοικίας των νοικοκυριών κ.λπ.). Αντιθέτως, η άλλη μεταβλητή Ψ ονομάζεται εξαρτημένη μεταβλητή επειδή οι τιμές της δεν ελέγχονται από τον ερευνητή δηλαδή, οι τιμές της εξαρτώνται από τις τιμές της μεταβλητής (π.χ. η ζήτηση ενός προϊόντος, η κατανάλωση ρεύματος κ.λπ.). Με άλλα λόγια, ως εξαρτημένη μεταβλητή επιλέγεται αυτή στην οποία αντανακλάται το αποτέλεσμα των μεταβολών της Χ, ή αυτή που καθοδηγείται από την ανεξάρτητη μεταβλητή Χ.

Με την ανάλυση απλής παλινδρόμησης, σκοπός είναι να βρούμε ένα υπόδειγμα, δηλαδή μια μαθηματική σχέση (ευθύγραμμη, πολυωνυμική, καμπυλόγραμμη, εκθετική κ.λπ.) η οποία να εκφράζει την εξάρτηση της μεταβλητής Υ από τη Χ. Για παράδειγμα, αν μια επιχείρηση για την παραγωγή ενός προϊόντος της, θέλει να γνωρίζει πως η μέση τιμή διάθεσής του από τις ανταγωνίστριες επιχειρήσεις επηρεάζει τη ζήτηση του προϊόντος

της, μπορούμε με βάση τη μαθηματική σχέση που θα εξειδικεύσουμε να προσδιορίσουμε (ή ακόμα και να προβλέψουμε) τη ζήτηση όταν η μέση τιμή διάθεσης του ανταγωνισμού είναι Χ ευρώ. Η ανάλυση απλής παλινδρόμησης που ακολουθεί, βασίζεται στην απλή υπόθεση ότι η εξαρτημένη μεταβλητή Υ μπορεί να προσεγγιστεί ικανοποιητικά από μια γραμμική συνάρτηση του Χ(Υ=α+βΧ).

Για τη δημιουργία ενός υποδείγματος απλής παλινδρόμησης, κάνουμε τις εξής υποθέσεις:

1. Η μεταβλητή Χ είναι ελεγχόμενη, δηλαδή γνωρίζουμε τις τιμές της. Αυτό σημαίνει ότι είναι μαθηματική μεταβλητή και όχι τυχαία.

2. Οι τιμές της μεταβλητής Χ υπολογίζονται χωρίς κάποιο σφάλμα.

3. Σε κάθε τιμή X_i της Χ αντιστοιχεί μια κατανομή πιθανότητας της τυχαίας μεταβλητής Υ και οι παρατηρηθείσες τιμές της Υ για δεδομένη τιμή X_i της Χ αποτελούν τυχαίο δείγμα που πάρθηκε από την κατανομή αυτή.

4. Οι διακυμάνσεις των κατανομών της τυχαίας μεταβλητής Υ είναι όλες ίσες μεταξύ τους.

5. Όλοι οι μέσοι των κατανομών της τυχαίας μεταβλητής Υ βρίσκονται πάνω στην ίδια ευθεία γραμμή (υπόθεση γραμμικότητας). Δηλαδή, για τον πληθυσμό ισχύει η ακόλουθη σχέση η οποία ονομάζεται απλή γραμμική παλινδρόμηση του πληθυσμού ή πληθυσμιακή ευθεία παλινδρόμησης:

$E(Y/X) = \alpha + \beta X$ ή $Y = \alpha + \beta X$	**(10.1)**

όπου:
Ε(Υ/Χ)= μέσος της κατανομής Υ για δεδομένη τιμή X_i της Χ.
α,β= συντελεστές παλινδρόμησης του πληθυσμού.

6. Όλες οι τιμές της Υ είναι ανεξάρτητες μεταξύ τους.

Επειδή στην πράξη συμβαίνει να παρατηρούνται σφάλματα κατά τον υπολογισμό των τιμών της μεταβλητής Υ, στην εξίσωση (10.1) θα πρέπει να προσθέσουμε τον όρο ε ο οποίος για δεδομένη τιμή X_i της Χ να σημειώνει τη διαφορά (απόκλιση) μεταξύ της παρατηρούμενης από τη θεωρητική (α+βΧ) τιμή της Υ. Δηλαδή,

$\varepsilon = Y - (\alpha + \beta X)$	**(10.2)**

Επομένως, προκύπτει το ακόλουθο στοχαστικό υπόδειγμα:

$Y = \alpha + \beta X + \varepsilon$	**(10.3)**

Στο παραπάνω υπόδειγμα, ο όρος ε ονομάζεται σφάλμα παλινδρόμησης ή διαταρακτικός όρος ή κατάλοιπο.

Οι υποθέσεις του γραμμικού υποδείγματος που σχετίζονται με το σφάλμα παλινδρόμησης είναι:

1. Για κάθε τιμή X_i της X, το σφάλμα παλινδρόμησης είναι μία τυχαία μεταβλητή η οποία έχει μέση τιμή μηδέν και διασπορά σταθερή η οποία δεν εξαρτάται από τη X.

2. Τα δειγματικά σφάλματα δεν σχετίζονται μεταξύ τους (η παραβίαση αυτής της υπόθεσης οδηγεί στο πρόβλημα της αυτοσυσχέτισης του διαταρακτικού όρου).

3. Κάθε δειγματικό σφάλμα, κατανέμεται με την ίδια διακύμανση (η παραβίαση αυτής της υπόθεσης οδηγεί στο πρόβλημα της ετεροσκεδαστικότητας).

4. Κάθε δειγματικό σφάλμα κατανέμεται κανονικά.

Στη συνέχεια, με τη μέθοδο των ελαχίστων τετραγώνων θα εκτιμήσουμε τις παραμέτρους του υποδείγματος της πληθυσμιακής παλινδρόμησης, δηλαδή τους συντελεστές α και β, ώστε η ευθεία γραμμή που θα προκύψει $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ (όπου $\hat{\alpha}$ και $\hat{\beta}$ εκτιμήτριες των α και β αντίστοιχα), να περιγράφει κατά τον καλύτερο δυνατό τρόπο τη σχέση μεταξύ των μεταβλητών X και Y.

10.2.1 Εκτίμηση των παραμέτρων α και β

Η σημειακή εκτίμηση $\hat{\alpha}$ και $\hat{\beta}$ των παραμέτρων α και β της πληθυσμιακής ευθείας παλινδρόμησης $Y = \alpha + \beta X$, επιτυγχάνεται με τη μέθοδο των ελαχίστων τετραγώνων, ελαχιστοποιώντας το άθροισμα των σφαλμάτων/καταλοίπων στο τετράγωνο, δηλαδή:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (10.4)$$

Οι σημειακές εκτιμήσεις $\hat{\alpha}$ και $\hat{\beta}$ των παραμέτρων (α και β) της πληθυσμιακής ευθείας παλινδρόμησης, οι οποίες δίνονται από τις σχέσεις:

$$\hat{\alpha} = \frac{1}{n}\sum_{i}^{n}(Y_i - \hat{\beta}X_i) = \overline{Y} - \hat{\beta}\overline{X} \quad (10.5)$$

$$\hat{\beta} = \frac{\sum_{i}^{n} X_iY_i - n\overline{X}\overline{Y}}{\sum_{i}^{n} X_i^2 - n\overline{X}^2} \text{ ή } \hat{\beta} = \frac{n\sum_{i}^{n} X_iY_i - \sum_{i}^{n} X_i \sum_{i}^{n} Y_i}{n\sum_{i}^{n} X_i^2 - (\sum_{i}^{n} X_i)^2} \text{ ή } \hat{\beta} = \frac{\sum_{i}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \quad (10.6)$$

Επομένως, η εξίσωση της ευθείας ελαχίστων τετραγώνων ή εκτιμηθέν υπόδειγμα (estimated model):

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$	**(10.7)**

χρησιμοποιείται για την εκτίμηση της πληθυσμιακής ευθείας παλινδρόμησης $Y = \alpha + \beta X$.

Ο εκτιμητής $\hat{\alpha}$ είναι ο σταθερός όρος της ευθείας και ερμηνεύεται ως η εκτίμηση της Y όταν X=0. Αντιθέτως, το $\hat{\beta}$ είναι η κλίση της ευθείας ελαχίστων τετραγώνων και υποδηλώνει την εκτιμώμενη μεταβολή της Y αν το X αυξηθεί κατά μια μονάδα.

Οι εκτιμητές $\hat{\alpha}$ και $\hat{\beta}$ που προέκυψαν με τη μέθοδο των ελαχίστων τετραγώνων, έχουν την ελάχιστη διακύμανση και ονομάζονται άριστοι αμερόληπτοι γραμμικοί εκτιμητές.

Παράδειγμα 10.1

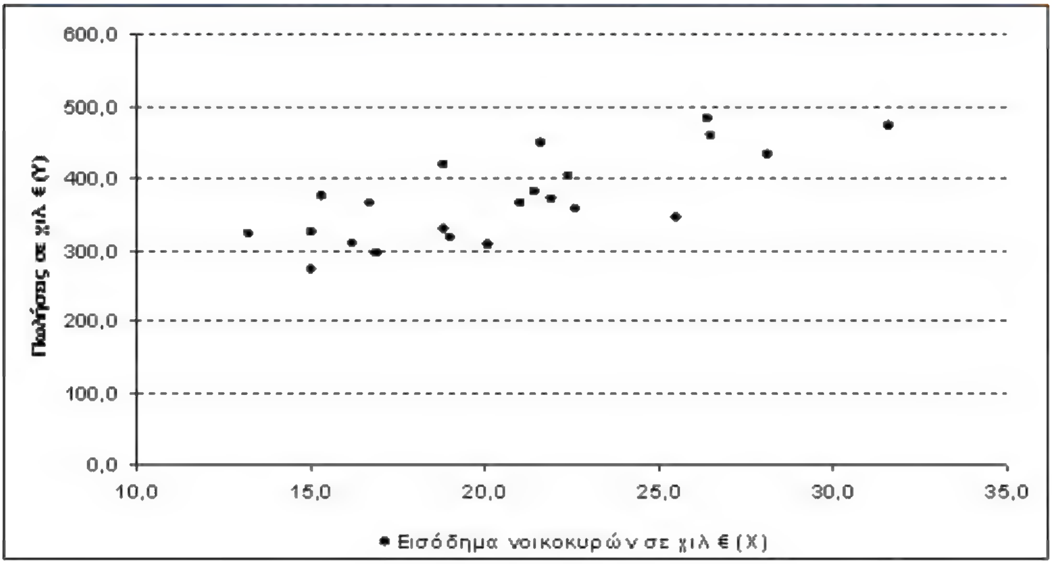
Επιχείρηση που δραστηριοποιείται στον χώρο της εστίασης, διαθέτει 22 υποκαταστήματα οικογενειακών εστιατορίων σε διάφορες περιοχές της Αθήνας. Η διοίκηση της εταιρείας, αποφάσισε να λειτουργήσει από το επόμενο έτος ακόμα ένα εστιατόριο και για το σκοπό αυτό, σας ανέθεσε να προσδιορίσετε την καλύτερη θέση για την ίδρυση και λειτουργία του. Μια έρευνα αγοράς δείχνει ότι για την επιλογή της κατάλληλης θέσης, θα πρέπει να ληφθούν υπόψη τρεις μεταβλητές, α) το μέσο οικογενειακό εισόδημα των κατοίκων των περιοχών δραστηριότητας των υποκαταστημάτων, β) ο αριθμός των κατοίκων που ζουν σε ακτίνα 3 χιλιομέτρων από τα υποκαταστήματά της και γ) ο ανταγωνισμός, δηλαδή το πλήθος των καταστημάτων άλλων εταιρειών που προσφέρουν ίδιες υπηρεσίες (υποκατάστατα ξένα καταστήματα) και δραστηριοποιούνται σε ακτίνα ενός χιλιομέτρου από τη θέση των υποκαταστημάτων της. Σε πρώτη φάση, ζητείται να εκτιμηθεί η γραμμική αιτιώδης σχέση μεταξύ πωλήσεων και εισοδήματος.

Στον πίνακα που ακολουθεί, παρουσιάζονται οι πωλήσεις των 22 υποκαταστημάτων της επιχείρησης καθώς και το μέσο οικογενειακό εισόδημα των κατοίκων της περιοχής δραστηριότητάς τους.

Απάντηση:

Κατ' αρχάς, από το επόμενο διάγραμμα διασποράς μεταξύ των πωλήσεων και του οικογενειακού εισοδήματος, υπάρχει μια εμφανής σχέση μεταξύ των δύο αυτών μεταβλητών. Δηλαδή, όσο αυξάνει το οικογενειακό εισόδημα, τόσο αυξάνουν οι πωλήσεις της επιχείρησης. Επίσης, παρατηρείται ότι οι πωλήσεις μεταβάλλονται συστηματικά γύρω από μια ευθεία γραμμή. Επομένως, μπορούμε να εκτιμήσουμε ένα γραμμικό υπόδειγμα της μορφής $Y = \alpha + \beta X$ στα δεδομένα του προβλήματος.

Διάγραμμα 10.1 Διάγραμμα διασποράς μεταξύ πωλήσεων και εισοδήματος



Πίνακας 10.1 Πωλήσεις και οικογενειακό εισόδημα

Πωλήσεις σε χιλ. € (Y)	Εισόδημα νοικοκυρών σε χιλ € (X)	YX	Y^2	X^2
323,8	13,2	4.274,2	104.846,4	174,2
356,6	22,6	8.059,2	127.163,6	510,8
295,7	16,9	4.997,3	87.438,5	285,6
366,0	21,0	7.686,0	133.956,0	441,0
448,5	21,6	9.687,6	201.152,3	466,6
273,8	15,0	4.107,0	74.966,4	225,0
370,7	21,9	8.118,3	137.418,5	479,6
482,8	26,4	12.745,9	233.095,8	697,0
345,8	25,5	8.817,9	119.577,6	650,3
324,2	15,0	4.863,0	105.105,6	225,0
434,4	28,1	12.206,6	188.703,4	789,6
473,6	31,6	14.965,8	224.297,0	998,6
316,7	19,0	6.017,3	100.298,9	361,0
307,7	20,1	6.184,8	94.679,3	404,0
310,0	16,2	5.022,0	96.100,0	262,4
381,1	21,4	8.155,5	145.237,2	458,0
420,0	18,8	7.896,0	176.400,0	353,4
376,0	15,3	5.752,8	141.376,0	234,1
365,6	16,7	6.105,5	133.663,4	278,9
403,8	22,4	9.045,1	163.054,4	501,8
458,8	26,5	12.158,2	210.497,4	702,3
328,9	18,8	6.183,3	108.175,2	353,4
$\sum Y = 8.164,5$	$\sum X = 454,0$	$\sum YX = 173.049,4$	$\sum Y^2 = 3.107.203,0$	$\sum X^2 = 9.852,4$
$\bar{Y} = 371,1$	$\bar{X} = 20,6$			

Χρησιμοποιώντας τα αποτελέσματα του Πίνακα 10.1 μπορούμε να εκτιμήσουμε την εξίσωση παλινδρόμησης των πωλήσεων (Y) ως προς το οικογενειακό εισόδημα (X). Επομένως:

$$\hat{\beta} = \frac{n\sum_i^n X_i Y_i - \sum_i^n X_i \sum_i^n Y_i}{n\sum_i^n X_i^2 - \sum_i^n (X_i)^2} = \frac{22 \cdot 173.049,4 \ - \ 450,0 \cdot 8.164,5}{229.852,4 \ - \ (454,0)^2} = 9,44$$

και $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 371,1 - 9,44 \cdot 454,0 = 176,3$

Άρα, η εξίσωση παλινδρόμησης είναι:

$$\hat{Y} = 176,3 + 9,44X$$

Από την εκτιμημένη εξίσωση παλινδρόμησης προκύπτει ότι αν το οικογενειακό εισόδημα των νοικοκυριών δεν αυξηθεί καθόλου, τότε οι μέσες πωλήσεις της επιχείρησης αναμένεται να είναι της τάξης των 176,3 χιλιάδων €. Αντιθέτως, αν το οικογενειακό εισόδημα αυξηθεί κατά χίλια €, τότε οι πωλήσεις θα αυξηθούν κατά 9,44 χιλ. €.

Στη παραπάνω εξίσωση παλινδρόμησης, ο συντελεστής $\hat{\beta}$ (η κλίση της ευθείας) μετράει την απόλυτη μεταβολή που επέρχεται στη μεταβλητή Y όταν η μεταβλητή X αυξηθεί ή μειωθεί κατά μία μονάδα. Χρησιμοποιώντας την έννοια της ελαστικότητας της μεταβλητής Y ως προς τη μεταβλητή X, μπορούμε να μετρήσουμε την ποσοστιαία μεταβολή της Y αν η X μεταβληθεί κατά 1%. Επομένως, για την περίπτωσή μας, η ελαστικότητας της μεταβλητής Y ως προς τη μεταβλητή X είναι:

$$e_{YX} = \hat{\beta} \cdot \frac{\bar{X}}{\bar{Y}} = 9,44 \cdot \frac{20,6}{371,1} = 0,52$$

Αυτό σημαίνει ότι, αν αυξηθεί το οικογενειακό εισόδημα κατά 1%, τότε οι πωλήσεις της επιχείρησης θα αυξηθούν κατά 0,52%.

10.2.2 Σφάλματα εκτίμησης ή κατάλοιπα

Η ευθεία ελαχίστων τετραγώνων όπως αυτή προέκυψε, αποτελεί μια προσεγγιστική σχέση σχετικά με το πώς συνδέονται οι τιμές X_i και Y_i των δεδομένων μας. Αν στην εξίσωση της ευθείας αντικαταστήσουμε όπου X τις τιμές X_i των δεδομένων, τότε παίρνουμε τις εκτιμήσεις $\hat{Y}_i$ των τιμών της Y.

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i,\ i=1,\dots,n$	**(10.8)**

οι οποίες θα πρέπει να βρίσκονται πάνω ή και κοντά στην ευθεία.

Οι διαφορές που προκύπτουν μεταξύ των εκτιμήσεων $\hat{Y}_i$ και των παρατηρήσεων Y_i, συμβολίζονται με $\hat{\varepsilon}_i$ και ονομάζονται σφάλματα εκτίμησης (errors of estimation) ή κατάλοιπα (residuals). Δηλαδή έχουμε:

$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i,\ i=1,\dots,n$	**(10.9)**

Το σφάλμα $\hat{\varepsilon}_i$ είναι μέρος της παρατήρησης Y_i που δεν συλλαμβάνεται από την ευθεία των ελαχίστων τετραγώνων και παίζει σημαντικό ρόλο κατά την αξιολόγηση του εκτιμηθέντος υποδείγματος.

Παράδειγμα 10.2

Με τα δεδομένα του παραδείγματος 10.1 να υπολογιστούν, οι θεωρητικές τιμές $\hat{Y}_i$ της μεταβλητής Y_i, τα κατάλοιπα καθώς και η γραφική παράσταση αυτών ως προς τις τιμές της X_i.

Απάντηση

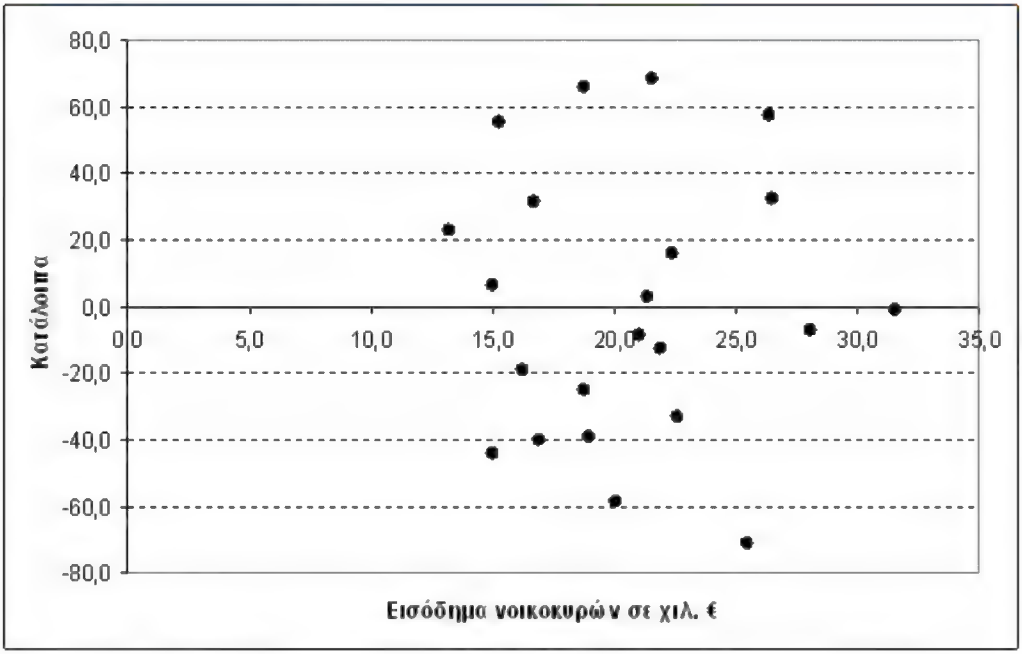
Οι θεωρητικές τιμές $\hat{Y}_i$ και των παρατηρήσεων Y_i υπολογίζονται με βάση την εξίσωση παλινδρόμησης $\hat{Y}=176,3+9,44X$, αν αντικαταστήσουμε σ' αυτήν τις τιμές X_i της μεταβλητής X.

Πίνακας 10.2 Θεωρητικές τιμές και κατάλοιπα παλινδρόμησης

Πωλήσεις σε χιλ. € (Y)	Εισόδημα νοικοκυρών σε χιλ € (X)	$\hat{Y}_i$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
323,8	13,2	300,9	22,9	523,2
356,6	22,6	389,6	-33,0	1.092,1
295,7	16,9	335,8	-40,1	1.611,9
366,0	21,0	374,5	-8,5	73,0
448,5	21,6	380,2	68,3	4.663,7
273,8	15,0	317,9	-44,1	1.946,1
370,7	21,9	383,0	-12,3	152,3
482,8	26,4	425,5	57,3	3.281,8
345,8	25,5	417,0	-71,2	5.072,1
324,2	15,0	317,9	6,3	39,5
434,4	28,1	441,6	-7,2	51,2
473,6	31,6	474,6	-1,0	1,0
316,7	19,0	355,7	-39,0	1.518,6
307,7	20,1	366,1	-58,4	3.404,9
310,0	16,2	329,2	-19,2	370,2
381,1	21,4	378,3	2,8	7,7
420,0	18,8	353,8	66,2	4.384,9
376,0	15,3	320,7	55,3	3.052,9
365,6	16,7	334,0	31,6	1.001,1
403,8	22,4	387,8	16,0	257,3
458,8	26,5	426,5	32,3	1.046,1
328,9	18,8	353,8	-24,9	619,1
				$\sum (Y_i - \hat{Y}_i)^2 = 34.170,7$

Από το διάγραμμα των καταλοίπων ως προς τις τιμές X_i, φαίνεται ότι τα κατάλοιπα κατανέμονται τυχαία γύρω από το μηδέν χωρίς καμία συστηματικότητα.

Διάγραμμα 10.2 Διάγραμμα καταλοίπων



10.2.3 Τυπικό σφάλμα εκτίμησης

Όπως έχουμε ήδη αναφέρει, η εξίσωση παλινδρόμησης χρησιμεύει για την εκτίμηση της εξαρτημένης μεταβλητής όταν δίνεται η τιμή της ανεξάρτητης. Αν και με την ευθεία παλινδρόμησης $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ παίρνουμε μια άριστη εκτίμηση της μέσης τιμής της εξαρτημένης μεταβλητής Y (αφού $\hat{\alpha}$ και $\hat{\beta}$ είναι αμερόληπτοι γραμμικοί εκτιμητές), εντούτοις θα πρέπει να γνωρίζουμε πόσο καλά παριστά αυτή τα δεδομένα μας ή κατά πόσο αυτή προσαρμόζει τη θεωρητική στην πραγματική κατάσταση.

Αν χρησιμοποιήσουμε ως μέσο αξιοπιστίας τη μέση απόκλιση τετραγώνου των τιμών Y_i της Y από τις εκτιμηθέντες τιμές $\hat{Y}_i$, δηλαδή τη διακύμανση των καταλοίπων γύρω από τη γραμμή παλινδρόμησης, τότε έχουμε:

$s^2_{Y/X} = \dfrac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-2} = \dfrac{\sum_i^n (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2}{n-2}$	**(10.10)**

Το $s^2_{Y/X}$, δίνει μια εκτίμηση της διασποράς των τυχαίων σφαλμάτων και ονομάζεται μέσο τετραγωνικό σφάλμα (mean squared error). Η θετική τετραγωνική του ρίζα της διασποράς των τυχαίων σφαλμάτων ($\sqrt{s^2_{Y/X}}$), ονομάζεται τυπικό σφάλμα εκτίμησης (standard error of estimate) ή τυπική απόκλιση των καταλοίπων (residual standard deviation).

Παράδειγμα 10.3

Με τα δεδομένα του παραδείγματος 10.1 να υπολογιστεί το τυπικό σφάλμα εκτίμησης.

Απάντηση:

Για να υπολογίσουμε το σφάλμα εκτίμησης θα πρέπει πρώτα να υπολογίσουμε την διακύμανση $s^2_{Y/X}$. Από τον πίνακα 10.2 προκύπτει ότι $\sum\left(Y_i - \hat{Y}_i\right)^2 = 34.170{,}7$,

επομένως $$s^2_{Y/X} = \frac{\sum_i^n \left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{34.170{,}7}{22\text{-}2} = 1.708{,}5$$.

Άρα, το σφάλμα εκτίμησης θα είναι: $$s_{Y/X} = \sqrt{s^2_{Y/X}} = \sqrt{1.708{,}5} = 41{,}33$$ και σημαίνει ότι οι τιμές X_i της X, διαφέρουν/αποκλίνουν από τις θεωρητικές τιμές $\hat{Y}_i$, κατά μέσο όρο κατά 41,33 χιλ. €. Επειδή η τιμή της είναι σχετικά μικρή, μπορούμε να πούμε ότι η ευθεία γραμμή προσαρμόζεται αρκετά καλά στα εμπειρικά δεδομένα.

10.2.4 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων της παραμέτρου β

Γνωρίζουμε ότι το $\hat{\beta}$ εκτιμά την κλίση β της ευθείας παλινδρόμησης στον πληθυσμό. Αυτό που μας ενδιαφέρει στη περίπτωση αυτή είναι, αφενός μεν να υπολογίσουμε το διάστημα στο οποίο θα βρίσκεται η κλίση β της ευθείας παλινδρόμησης στον πληθυσμό (με ένα ορισμένο επίπεδο σημαντικότητας 1-α) και αφετέρου, να ελέγξουμε τη μηδενική υπόθεση ότι το β ισούται με το μηδέν, με αποτέλεσμα η παρατηρούμενη διαφορά του $\hat{\beta}$ από το β να οφείλεται στις διακυμάνσεις της δειγματοληψίας.

Διάστημα εμπιστοσύνης

α) Όταν γνωρίζουμε την πληθυσμιακή διακύμανση $\sigma^2_{Y/X}$, τότε:

$z=\dfrac{\hat{\beta}-\beta}{\sqrt{\sigma^2_{\hat{\beta}}}}=\dfrac{\hat{\beta}-\beta}{\sqrt{\dfrac{\sigma^2_{Y/X}}{\sum_i^n\left(X_i-\overline{X}\right)^2}}}=\dfrac{\hat{\beta}-\beta}{\dfrac{\sigma_{Y/X}}{\sqrt{\sum_i^n\left(X_i-\overline{X}\right)^2}}}$	**(10.11)**

Το $\sigma_{Y/X}$ είναι η τετραγωνική ρίζα του $\sigma^2_{Y/X}$ και μπορεί να εκτιμηθεί από την παραπάνω σχέση. Επομένως, το διάστημα εμπιστοσύνης του $\sigma^2_{Y/X}$ είναι γνωστό.

β) Όταν το δείγμα των παρατηρήσεων είναι μεγάλο (n>30) και δεν γνωρίζουμε την πληθυσμιακή διακύμανση $\sigma^2_{Y/X}$, τότε ακολουθούμε την ίδια διαδικασία της παραπάνω περίπτωσης, με τη διαφορά ότι θα πρέπει να χρησιμοποιήσουμε μια εκτίμηση της. Μια αμερόληπτη εκτίμηση της $\sigma^2_{Y/X}$ δίνεται από τον εκτιμητή:

$s^2_{Y/X}=\dfrac{\sum_i^n\left(Y_i-\hat{Y}_i\right)^2}{n-2}$	**(10.12)**

γ) Όταν το δείγμα των παρατηρήσεων είναι μικρό (n<30) και δεν γνωρίζουμε την πληθυσμιακή διακύμανση $\sigma^2_{Y/X}$, τότε $t=\dfrac{(\hat{\beta}-\beta)}{s_{\hat{\beta}}}$.. Το $s_{\hat{\beta}}$ ονομάζεται τυπικό σφάλμα (standard error) του συντελεστή $\hat{\beta}$ και δίνεται από τη σχέση:

$s_{\hat{\beta}}=\dfrac{s_{Y/X}}{\sqrt{\sum_i^n\left(X_i-\overline{X}\right)^2}}$	**(10.13)**

Επομένως, το διάστημα στο οποίο θα βρίσκεται η κλίση β (σε επίπεδο σημαντικότητας α) είναι:

$\beta = \hat{\beta} \pm t_{n-2,\alpha/2} s_{\hat{\beta}}$ ή $\hat{\beta} - t_{n-2,\alpha/2} s_{\hat{\beta}} \leq \beta \leq \hat{\beta} + t_{n-2,\alpha/2} s_{\hat{\beta}}$	**(10.14)**

Έλεγχος υποθέσεων

Όταν ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η παράμετρος β της ευθείας παλινδρόμησης στον πληθυσμό είναι β=0, δηλαδή ότι οι μεταβλητές X και Y είναι ανεξάρτητες, τότε ο έλεγχος διατυπώνεται ως εξής:

H_0: $\beta = 0$

H_r: i) $\beta \neq 0$

ii) $\beta > 0$

iii) $\beta < 0$

Η στατιστική ελέγχου που χρησιμοποιείται σ' αυτή την περίπτωση είναι:

$t = \dfrac{\hat{\beta} - \beta}{s_{\hat{\beta}}} = \dfrac{\hat{\beta}}{s_{\hat{\beta}}} = \dfrac{\hat{\beta}}{s_{Y/X}} \sqrt{\sum_i^n (X_i - \overline{X})^2} = \dfrac{\hat{\beta} s_X}{s_{Y/X}} \sqrt{(n-1)}$	**(10.15)**

Επομένως, το κριτήριο απόφασης για τις παραπάνω τρεις υποθέσεις ελέγχου, είναι αντίστοιχα:

- Αν $t = \dfrac{\hat{\beta} - \beta}{s_{\hat{\beta}}} < t_{n-2,\alpha/2}$ ή $t = \dfrac{\hat{\beta} - \beta}{s_{\hat{\beta}}} > t_{n-2,\alpha/2}$, τότε απορρίπτουμε τη μηδενική υπόθεση

σε επίπεδο σημαντικότητας α και για n-2 βαθμούς ελευθερίας και συμπεραίνουμε ότι η παρατηρούμενη διαφορά του $\hat{\beta}$ από το μηδέν είναι στατιστικά σημαντική.

- Αν $t = \dfrac{\hat{\beta} - \beta}{s_{\hat{\beta}}} > t_{n-2,\alpha}$, τότε απορρίπτουμε τη μηδενική υπόθεση.

- Αν $t = \dfrac{\hat{\beta} - \beta}{s_{\hat{\beta}}} < t_{n-2,\alpha}$, τότε απορρίπτουμε τη μηδενική υπόθεση.

Τέλος, τη στατιστική σημαντικότητα του $\hat{\beta}$ και των αντίστοιχων μεταβλητών, μπορούμε να την ελέγξουμε εμπειρικά με το κριτήριο $s_{\hat{\beta}} < 0,5\hat{\beta}$.. Δηλαδή, αν το τυπικό σφάλμα είναι μικρότερο από το μισό της εκτιμημένης τιμής $\hat{\beta}$, τότε η παράμετρος αυτή (ή η κλίση της ευθείας) είναι στατιστικά σημαντική.

Παράδειγμα 10.4

Με τα δεδομένα του παραδείγματος 10.1, να ελεγχθεί η στατιστική σημαντικότητα του συντελεστή παλινδρόμησης $\hat{\beta}$ σε επίπεδο σημαντικότητας α = 0,05 δηλαδή, ότι οι μεταβλητές X και Y είναι ανεξάρτητες και να προσδιοριστεί διάστημα εμπιστοσύνης του πληθυσμιακού συντελεστή παλινδρόμησης β με πιθανότητα 95%.

Απάντηση:

α) Για τον έλεγχο της ανεξαρτησίας των μεταβλητών X και Y, υποθέτουμε ότι ο πληθυσμιακός συντελεστής παλινδρόμησης είναι μηδέν. Επομένως, ο έλεγχος διατυπώνεται ως εξής:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

Επειδή το δείγμα είναι μικρό (N=22<30), κατά τον έλεγχο θα χρησιμοποιήσουμε τη στατιστική t-student. Για να προσδιορίσουμε την τιμή του κριτηρίου t, θα πρέπει κατ' αρχάς να υπολογίσουμε το τυπικό σφάλμα εκτίμησης του $\hat{\beta}$:

$$s_{\hat{\beta}} = \frac{s_{Y/X}}{\sqrt{\sum_i^n (X_i - \overline{X})^2}} = \frac{41,33}{\sqrt{483,53}} = 1,88$$

Όπου $\sum_i^n (X_i - \overline{X})^2 = \sum_i^n X_i^2 - \frac{\left(\sum_i^n X_i\right)^2}{n} = 9.852,4 - \frac{(454,0)^2}{22} = 483,53$

Επομένως, $t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}} = \frac{9,44}{1,88} = 5,02$

Από τον Πίνακα κριτικών τιμών της κατανομής t-student και για α=0,05 και ν = 22-2 = 20, βρίσκουμε ότι $t_{22,0,05}$ =2,074.

Επειδή t=5,02> $t_{22,0,05}$ =2,074 , απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική, που σημαίνει ότι υπάρχει εξάρτηση μεταξύ των μεταβλητών X και Y, δηλαδή μεταξύ του εισοδήματος των νοικοκυριών και των πωλήσεων της επιχείρησης.

β) Το διάστημα εμπιστοσύνης του πληθυσμιακού συντελεστή παλινδρόμησης β με πιθανότητα 95% (ή επίπεδο σημαντικότητας α = 1-0,95=0,05), προσδιορίζεται από την ακόλουθη διπλή ανισότητα:

$$\hat{\beta}-t_{n-2,\alpha/2}s_{\hat{\beta}} \leq \beta \leq \hat{\beta}+t_{n-2,\alpha/2}s_{\hat{\beta}} \text{ ή}$$

$$9{,}44-2{,}075\cdot 1{,}88 \leq \beta \leq 9{,}44+2{,}075\cdot 1{,}88 \text{ ή}$$

$$5{,}54 \leq \beta \leq 13{,}34$$

Επομένως, η τιμή του συντελεστή παλινδρόμησης στον πληθυσμό, κυμαίνεται από 5,34 έως 13,34.

10.2.5 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων της παραμέτρου

Η διατύπωση ενός διαστήματος εμπιστοσύνης για το α και ενός ελέγχου για να διαπιστώσουμε ότι το α είναι ίσο με μια συγκεκριμένη τιμή, μπορεί να γίνει με παρόμοιο τρόπο μ' αυτόν που περιγράψαμε παραπάνω για το β.

Όπως και στην προηγούμενη περίπτωση, έτσι και εδώ, η σταθερά α έχει διακύμανση την:

$$\sigma^2_{\hat{\alpha}} = \frac{\sigma^2_{Y/X}\sum_{i}^{n} X_i^2}{n\sum_{i}^{n}(X_i-\overline{X})^2} \qquad \textbf{(10.16)}$$

Η τετραγωνική ρίζα της διακυμάνσης του $\hat{\alpha}$ ονομάζεται τυπικό σφάλμα (standard eror) και συμβολίζεται με $s_{\hat{\alpha}}$.

Διάστημα εμπιστοσύνης

α) Όταν γνωρίζουμε την πληθυσμιακή διακύμανση $\sigma^2_{Y/X}$, τότε

$$z=\frac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}}=\frac{\hat{\alpha}-\alpha}{\sqrt{\sigma^2_{\hat{\alpha}}}}=\frac{\hat{\alpha}-\alpha}{\sqrt{\dfrac{\sigma^2_{Y/X}\sum_{i}^{n} X_i^2}{n\sum_{i}^{n}(X_i-\overline{X})^2}}}=\frac{\hat{\alpha}-\alpha}{s_{Y/X}\sqrt{\dfrac{\sum_{i}^{n} X_i^2}{n\sum_{i}^{n}(X_i-\overline{X})^2}}} \qquad \textbf{(10.17)}$$

και το διάστημα εμπιστοσύνης του είναι γνωστό.

β) Όταν το δείγμα των παρατηρήσεων είναι μεγάλο (n>30) και δεν γνωρίζουμε την πληθυσμιακή διακύμανση $\sigma^2_{Y/X}$, τότε ακολουθούμε την παραπάνω διαδικασία, με τη διαφορά ότι θα πρέπει να χρησιμοποιήσουμε στη θέση της τον εκτιμητή $s^2_{Y/X}$..

γ) Όταν το δείγμα των παρατηρήσεων είναι μικρό (n<30) και δεν γνωρίζουμε την πληθυσμιακή διακύμανση $\sigma^2_{Y/X}$, τότε ο λόγος $\dfrac{(\hat{\alpha}-\alpha)}{s_{\hat{\alpha}}}$ ακολουθεί την κατανομή t-student με v=n-2 βαθμούς ελευθερίας, επομένως $t=\dfrac{(\hat{\alpha}-\alpha)}{s_{\hat{\alpha}}}$. .

Επομένως, το διάστημα στο οποίο θα βρίσκεται η τιμή του α (σε επίπεδο σημαντικότητας α) είναι:

$\alpha=\hat{\alpha}\pm t_{n-2,\alpha/2}s_{\hat{\alpha}}$ ή $\hat{\alpha}-t_{n-2,\alpha/2}s_{\hat{\alpha}}\leq\alpha\leq\hat{\alpha}+t_{n-2,\alpha/2}s_{\hat{\alpha}}$	**(10.18)**

Έλεγχος υποθέσεων

Όταν ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η παράμετρος β της ευθείας παλινδρόμησης στον πληθυσμό είναι β=0, δηλαδή ότι οι μεταβλητές X και Y είναι ανεξάρτητες, τότε ο έλεγχος διατυπώνεται ως εξής:

H_0: $\alpha=0$

H_1: i) $\alpha\neq 0$

ii) $\alpha>0$

iii) $\alpha<0$

Η στατιστική ελέγχου που χρησιμοποιείται σ' αυτή την περίπτωση είναι:

$t=\dfrac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}}$	**(10.19)**

Επομένως, το κριτήριο απόφασης για τις παραπάνω τρεις υποθέσεις ελέγχου, είναι αντίστοιχα:

- Αν $t=\dfrac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}}<t_{n-2,\alpha/2}$ ή $t=\dfrac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}}>t_{n-2,\alpha/2}$, τότε απορρίπτουμε τη μηδενική υπόθεση

σε επίπεδο σημαντικότητας α και για n-2 βαθμούς ελευθερίας.

- Αν $t=\dfrac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}}>t_{n-2,\alpha}$, τότε απορρίπτουμε τη μηδενική υπόθεση.

- Αν $t = \frac{\hat{\alpha} - \alpha}{s_{\hat{\alpha}}} < t_{n-2,\alpha}$, τότε απορρίπτουμε τη μηδενική υπόθεση.

Τέλος, τη στατιστική σημαντικότητα του $\hat{\alpha}$ και των αντίστοιχων μεταβλητών μπορούμε να την ελέγξουμε εμπειρικά με το κριτήριο $s_{\hat{\alpha}} < 0,5\hat{\alpha}$. Δηλαδή, αν το τυπικό σφάλμα είναι μικρότερο από το μισό της εκτιμημένης τιμής $\hat{\alpha}$, τότε η παράμετρος αυτή είναι στατιστικά σημαντική.

10.2.6 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων του μέσου της κατανομής Y ($\overline{Y}_X$)

Πρόκειται για την περίπτωση κατά την οποία θέλουμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης εντός του οποίου βρίσκεται (με δεδομένη πιθανότητα) η αληθής μέση τιμή της μεταβλητής Y.

Έστω η γραμμή παλινδρόμησης $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$. Η εκτίμηση του μέσου μιας κατανομής Y για μια δεδομένη τιμή X_k της X, η οποία σημειώνεται με $\hat{Y}_i$ (η τιμή της $\hat{Y}$ στη θέση k) είναι ίση με:

$\hat{\hat{Y}}_k = \hat{\alpha} + \hat{\beta}X_k$	**(10.20)**

Στην περίπτωση αυτή, ο εκτιμητής $\hat{\hat{Y}}_k$ έχει αμερόληπτη διακύμανση την:

$\sigma^2_{\hat{Y}_k} = \frac{\sigma^2_{Y/X}}{n} + \frac{(X_k - \overline{X})^2 \sigma^2_{Y/X}}{\sum_i^n (X_i - \overline{X})^2} = \sigma^2_{Y/X}\left[\frac{1}{n} + \frac{(X_k - \overline{X})^2}{\sum_i^n (X_i - \overline{X})^2}\right]$	**(10.21)**

Η τετραγωνική ρίζα της διακύμανσης $\sigma^2_{\hat{Y}_k}$, ονομάζεται τυπική απόκλιση της μέσης τιμής της Y για δεδομένο X και συμβολίζεται με $\sigma_{\hat{Y}_k}$.

Διάστημα εμπιστοσύνης

α) Όταν γνωρίζουμε τη διακύμανση $\sigma^2_{Y/X}$, τότε το διάστημα εμπιστοσύνης της για επίπεδο σημαντικότητας α είναι:

$\hat{Y}_k - z_{\alpha/2} \cdot \sigma_{\hat{Y}_k} < \bar{Y}_k < \hat{Y}_k + z_{\alpha/2} \cdot \sigma_{\hat{Y}_k}$	**(10.22)**

β) Όταν το δείγμα των παρατηρήσεων είναι μεγάλο (n>30) και δεν γνωρίζουμε τη διακύμανση $\sigma^2_{Y/X}$, τότε:

$s_{\hat{Y}} = \frac{s_{Y/X}}{n} + \frac{(X_k - \bar{X})^2 s_{Y/X}}{\sum_i^n (X_i - \bar{X})^2} = s_{Y/X}\sqrt{\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum_i^n (X_i - \bar{X})^2}}$	**(10.23)**

Επομένως, το διάστημα εμπιστοσύνης της $\bar{Y}_k$ για επίπεδο σημαντικότητας α είναι:

$\hat{Y}_k - z_{\alpha/2} \cdot s_{\hat{Y}} < \bar{Y}_k < \hat{Y}_k + z_{\alpha/2} \cdot s_{\hat{Y}}$	**(10.24)**

γ) Όταν το δείγμα των παρατηρήσεων είναι μικρό (N<30) και δεν γνωρίζουμε τη διακύμανση $\sigma^2_{Y/X}$, τότε ο λόγος $\frac{\hat{Y}_k - \bar{Y}_k}{S_{\bar{Y}}} = t$ ακολουθεί την κατανομή t-student v=n-2 με βαθμούς ελευθερίας.

Επομένως, το διάστημα στο οποίο θα βρίσκεται η $\bar{Y}_k$ σε επίπεδο σημαντικότητας α, είναι:

$\hat{Y}_k - t_{n,\alpha/2} \cdot s_{\hat{Y}} < \bar{Y}_k < \hat{Y}_k + t_{n,\alpha/2} \cdot s_{\hat{Y}}$	**(10.25)**

Έλεγχος υποθέσεων

Όταν για παράδειγμα ενδιαφερόμαστε να ελέγξουμε την υπόθεση:

$$H_0: \hat{Y} = Y_0$$
$$H_1: \hat{Y} \neq Y_0$$

σε επίπεδο σημαντικότητας α, για n<30 και για δεδομένη τιμή X_k της X, τότε χρησιμοποιούμε τη στατιστική $t = \dfrac{\hat{Y} - Y_0}{s_{\hat{Y}}}$.

Επομένως, απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας α και για n-2 βαθμούς ελευθερίας αν $t = \dfrac{\hat{Y} - Y_0}{s_{\hat{Y}}} < t_{n-2,\alpha/2}$ ή $t = \dfrac{\hat{Y} - Y_0}{s_{\hat{Y}}} > t_{n-2,\alpha/2}$.

Λοιποί έλεγχοι που αναφέρθηκαν παραπάνω, μπορούν να διατυπωθούν αναλόγως και σ' αυτή την περίπτωση.

Παράδειγμα 10.5

Με τα δεδομένα του παραδείγματος 10.1, να προσδιοριστεί διάστημα εμπιστοσύνης των πωλήσεων με πιθανότητα 95%, όταν το εισόδημα των νοικοκυριών είναι 21 χιλ.€.

Απάντηση:

Όταν το επίπεδο εισοδήματος των νοικοκυριών είναι 21,0 χιλ. € (δηλαδή X_k =21,0 χιλ. €), τότε η μέση αξία των πωλήσεων της επιχείρησης θα είναι:

$\hat{\hat{Y}}_{k=21,0} = 176,3 + 9,44 \cdot 21,0 = 374,5$ χιλ. €.

Επειδή το δείγμα είναι μικρό (n=2<30),θα χρησιμοποιήσουμε τη στατιστική t-student. Για να προσδιορίσουμε την τιμή του κριτηρίου t, θα πρέπει κατ' αρχάς να υπολογίσουμε το τυπικό σφάλμα εκτίμησης:

Από τα παραδείγματα 10.3 και 10.4 έχουμε υπολογίσει ότι:

$$s_{Y/X} = 41,33 \quad \text{και} \quad \sum_i^n (X_i - \bar{X})^2 = \sum_i^n X_i^2 - \frac{\left(\sum_i^n X_i\right)^2}{n} = 9.852,4 - \frac{(454,0)^2}{22} = 483,53$$

Επομένως, το τυπικό σφάλμα εκτίμησης είναι ίσο με:

$$s_{\hat{Y}} = s_{Y/X}\sqrt{\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum_i^n (X_i - \bar{X})^2}} = 41,33\sqrt{\frac{1}{22} + \frac{(21,0 - 20,6)^2}{483,53}} = 8,84$$

Από τον Πίνακα κριτικών τιμών της κατανομής t-student και για α=1-0.95=0,05 και ν = 22-2 = 20, βρίσκουμε ότι $t_{20,0.025}$ =2,086.

Μετά τα παραπάνω, το ζητούμενο διάστημα εμπιστοσύνης προσδιορίζεται από την ακόλουθη διπλή ανισότητα:

$$\hat{\bar{Y}}_k - t_{n,\alpha/2}\cdot s_{\hat{\bar{Y}}} < \bar{Y}_k < \hat{Y}_k + t_{n,\alpha/2}\cdot s_{\hat{\bar{Y}}} \quad \text{ή}$$

$$374{,}5 - 2{,}086\cdot 8{,}84 < \bar{Y}_k < 374{,}5 + 2{,}086\cdot 8{,}84 \quad \text{ή}$$

$$356{,}06 < \bar{Y}_k < 392{,}94$$

Επομένως, με εισόδημα νοικοκυριών 21,0 χιλ. €, οι μέσες πωλήσεις της επιχείρησης θα κυμαίνονται από 356,,06 χιλ. € έως 392,94 χιλ. €.

10.2.7 Εκτίμηση μιας τιμής της Υ (Πρόβλεψη)

Πολλές φορές επιθυμούμε να εκτιμήσουμε (προβλέψουμε) μια ατομική τιμή (παρατήρηση) της μεταβλητής Υ και όχι τη μέση τιμή αυτής όπως είδαμε παραπάνω. Η πραγματοποίηση έγκυρων προβλέψεων απαιτεί το εκτιμημένο υπόδειγμα γραμμής παλινδρόμησης $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ να έχει αξιολογηθεί και να έχει δώσει ικανοποιητικά αποτελέσματα. Η πρόβλεψη μιας τιμής της Υ η οποία αντιστοιχεί σε μια δεδομένη τιμή της X_p έστω και συμβολίζεται με $\hat{Y}_p$, βρίσκεται αν αντικαταστήσουμε την τιμή στο εκτιμημένο υπόδειγμα γραμμής παλινδρόμησης. Δηλαδή,

$\hat{Y}_p = \hat{\alpha} + \hat{\beta}X_p$	**(10.26)**

Στόχος στην περίπτωση αυτή, είναι να κατασκευάσουμε ένα διάστημα εμπιστοσύνης εντός του οποίου θα βρίσκεται (με δεδομένη πιθανότητα) η αληθής ατομική (προβλεπόμενη) τιμή ($\hat{Y}_p$) της μεταβλητής Υ (ανάλογα όπως και στη προηγούμενη περίπτωση της μέσης τιμής).

Στην περίπτωση αυτή, ο εκτιμητής $\hat{Y}_p$ έχει αμερόληπτη διακύμανση την:

$\sigma^2_{\hat{Y}} = \sigma^2_{Y/X} + \dfrac{\sigma^2_{Y/X}}{n} + \dfrac{(X_p - \bar{X})^2 \sigma^2_{Y/X}}{\sum_i^n (X_i - \bar{X})^2} = \sigma^2_{Y/X}\left[1 + \dfrac{1}{n} + \dfrac{(X_p - \bar{X})^2}{\sum_i^n (X_i - \bar{X})^2}\right]$	**(10.27)**

Η τετραγωνική ρίζα της διακύμανσης $\sigma^2_{\hat{Y}}$, ονομάζεται τυπική απόκλιση της ατομικής τιμής της Y για δεδομένο X και συμβολίζεται με $\sigma_{\hat{Y}}$.

Διάστημα εμπιστοσύνης

α) Όταν γνωρίζουμε τη διακύμανση $\sigma^2_{Y/X}$, τότε το διάστημα εμπιστοσύνης της $\hat{Y}_p$ για επίπεδο σημαντικότητας α είναι:

$\hat{Y}_p - z_{\alpha/2} \cdot \sigma_{\hat{Y}} < Y_p < \hat{Y}_p + z_{\alpha/2} \cdot \sigma_{\hat{Y}}$	**(10.28)**

β) Όταν το δείγμα των παρατηρήσεων είναι μεγάλο (n>30) και δεν γνωρίζουμε τη διακύμανση $\sigma^2_{Y/X}$, τότε η τυπική απόκλισή της $s^2_{\hat{Y}}$ που συμβολίζεται με $s^2_{\hat{Y}}$ δίνεται από τη σχέση:

$s_{\hat{Y}} = s_{Y/X} + \frac{s_{Y/X}}{n} + \frac{(X_p - \overline{X})^2 s_{Y/X}}{\sum_i^n (X_i - \overline{X})^2} = s_{Y/X} \sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_i^n (X_i - \overline{X})^2}}$	**(10.29)**

Επομένως, το διάστημα εμπιστοσύνης της Y_p για επίπεδο σημαντικότητας α είναι:

$\hat{Y}_p - z_{\alpha/2} \cdot s_{\hat{Y}} < Y_p < \hat{Y}_p + z_{\alpha/2} \cdot s_{\hat{Y}}$	**(10.30)**

γ) Όταν το δείγμα των παρατηρήσεων είναι μικρό (n<30) και δεν γνωρίζουμε τη διακύμανση $\sigma^2_{Y/X}$, τότε ο λόγος $\frac{\hat{Y}_p - Y_p}{S_{\hat{Y}}} = t$ ακολουθεί την κατανομή t-student με v=n-2 βαθμούς ελευθερίας.

Επομένως, το διάστημα εμπιστοσύνης (διάστημα προβλέψεων) σε επίπεδο σημαντικότητας α στο οποίο θα βρίσκεται η τιμή $\hat{Y}_p$ είναι:

$\hat{Y}_p - t_{n,\alpha/2} \cdot s_{\hat{Y}} < Y_p < \hat{Y}_p + t_{n,\alpha/2} \cdot s_{\hat{Y}}$	**(10.31)**

Ομοίως, όπως και στην προηγούμενη περίπτωση της μέσης τιμής της Y, μπορούμε να διατυπώσουμε ελέγχους υποθέσεων σχετικά με τη δεδομένη τιμή X_p της X.

Παράδειγμα 10.6

Με τα δεδομένα του παραδείγματος 10.1, να προσδιοριστεί διάστημα εμπιστοσύνης (όρια προβλέψεων) των πωλήσεων με πιθανότητα 95%, όταν το εισόδημα των νοικοκυριών προβλέπεται να ανέλθει στα 32 χιλ.€.

Απάντηση:

Αν και το ζητούμενο φαντάζει εξωπραγματικό, αφού το μέσο οικογενειακό εισόδημα είναι 21,6 χιλ. €, εντούτοις μπορούμε πρακτικά να προσδιορίσουμε τα όρια πρόβλεψης του προβλεπόμενου οικογενειακού εισοδήματος. Τα όρια πρόβλεψης αναμένονται περισσότερο διευρυμένα αφού επιλέξαμε προβλεπόμενο εισόδημα τέτοιο που απέχει κατά πολύ από το μέσο εισόδημα. Στη περίπτωση αυτή, ακολουθούμε την ίδια διαδικασία που αναφέραμε παραπάνω για την εκτίμηση του διαστήματος εμπιστοσύνης των πωλήσεων για δεδομένη τιμή του εισοδήματος.

Όταν το εισόδημα των νοικοκυρών προβλέπεται να ανέλθει στις 32,0 χιλ. € (δηλαδή X_p =32 χιλ. €), τότε οι προβλεπόμενες πωλήσεις της επιχείρησης θα είναι:

$$\hat{Y}_{p=32,0} = 176,3 + 9,44 \cdot 32,0 = 478,4 \text{ χιλ. €.}$$

Επειδή το δείγμα είναι μικρό (N=22<30), θα χρησιμοποιήσουμε τη στατιστική t-student. Για να προσδιορίσουμε την τιμή του κριτηρίου t, θα πρέπει κατ' αρχάς να υπολογίσουμε το τυπικό σφάλμα εκτίμησης:

Από τα παραδείγματα 10.3 και 10.4 έχουμε υπολογίσει ότι:

$$s_{Y/X} = 41,33 \quad \text{και} \quad \sum_i^n (X_i - \overline{X})^2 = \sum_i^n X_i^2 - \frac{\left(\sum_i^n X_i\right)^2}{n} = 9.852,4 - \frac{(454,0)^2}{22} = 483,53$$

Επομένως, το τυπικό σφάλμα εκτίμησης είναι ίσο με:

$$s_{\hat{Y}} = s_{Y/X}\sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_i^n (X_i - \overline{X})^2}} = 41,33\sqrt{1 + \frac{1}{22} + \frac{(32,0 - 20,6)^2}{483,53}} = 47,41$$

Από τον Πίνακα κριτικών τιμών της κατανομής t-student και για α=1-0.95=0,05 και ν = 22-2 = 20, βρίσκουμε ότι $t_{20,0.025} = 2,086$.

Μετά τα παραπάνω, το ζητούμενο όριο πρόβλεψης των πωλήσεων προσδιορίζεται από την ακόλουθη διπλή ανισότητα:

$$\hat{Y}_p - t_{n,\alpha/2} \cdot s_{\hat{Y}_p} < Y_p < \hat{Y}_p + t_{n,\alpha/2} \cdot s_{\hat{Y}} \quad \text{ή}$$

$$478{,}4 - 2{,}086 \cdot 47{,}41 < Y_p < 478{,}4 + 2{,}086 \cdot 47{,}41 \quad \text{ή} \quad 379{,}5 < Y_p < 577{,}3$$

Επομένως, με προβλεπόμενο εισόδημα νοικοκυριών 32,0 χιλ. €, οι προβλεπόμενες πωλήσεις της επιχείρησης θα κυμαίνονται από 379,5 χιλ. € έως 577,3 χιλ. €.

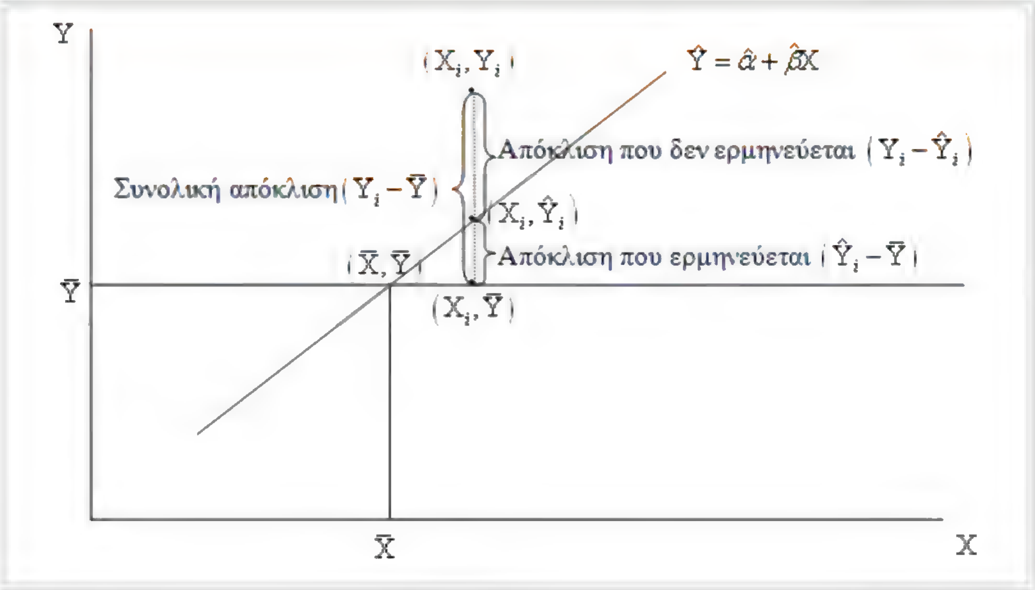
10.2.8 Ανάλυση διακύμανσης

Η ανάλυση διακύμανσης είναι μια κρίσιμη διαδικασία κατά την αξιολόγηση της εξίσωσης παλινδρόμησης που προσαρμόσαμε στα δεδομένα μας. Οι πληροφορίες που μας παρέχει, δεν εστιάζονται μόνο στο να διαπιστώσουμε αν είναι μικρή ή μεγάλη, τόσο η διασπορά των εκτιμημένων τιμών της Y γύρω από τη μέση τιμή $\bar{Y}$, όσο και η διασπορά των δειγματικών τιμών της Y γύρω από την εκτιμημένη ευθεία παλινδρόμησης, αλλά και να ελέγξουμε αφενός μεν την ύπαρξη ή όχι γραμμικής σχέσης μεταξύ των μεταβλητών Y, X και αφετέρου (όπως θα δούμε στην επόμενη ενότητα) το πόσο καλά η ευθεία των ελαχίστων τετραγώνων προσαρμόζεται πάνω στις δειγματικές τιμές.

Όπως έχει αναφερθεί, η ευθεία ελαχίστων τετραγώνων αφήνει το μικρότερο άθροισμα των τετραγώνων των καταλοίπων (σφαλμάτων) SSE. Για να δηλώνει όμως η τιμή αυτή καλή προσαρμογή ή όχι της ευθείας των ελαχίστων τετραγώνων στα δεδομένα, εξαρτάται από τη διακύμανση των παρατηρήσεων της Y_i.

Στο διάγραμμα που ακολουθεί παρουσιάζεται η ευθεία των ελαχίστων τετραγώνων καθώς και τρία σημεία (ζεύγη τιμών των μεταβλητών X_i και Y_i). Το ζεύγος τιμών (X_i, Y_i), το $(X_i, \hat{Y}_i)$ που βρίσκεται πάνω στην ευθεία των ελαχίστων τετραγώνων και το $(X_i, \bar{Y})$ που βρίσκεται πάνω στη γραμμή του μέσου της Y. Με βάση τα σημεία αυτά και τις ευθείες των ελαχίστων τετραγώνων και μέσης τιμής του Y, δημιουργούνται οι ακόλουθες τρεις διαφορές:

- Η διαφορά $(Y_i - \bar{Y})$, ονομάζεται συνολική απόκλιση της τιμής από το μέσο.
- Η διαφορά $(\hat{Y}_i - \bar{Y})$, ονομάζεται απόκλιση που ερμηνεύτηκε, αφού δείχνει κατά πόσο μειώθηκε η συνολική απόκλιση μετά την προσαρμογή στα στοιχεία της ευθείας παλινδρόμησης.
- Η διαφορά $(Y_i - \hat{Y}_i)$, ονομάζεται απόκλιση που δεν ερμηνεύτηκε, αφού δείχνει κατά πόσο δεν μειώθηκε η συνολική απόκλιση μετά την προσαρμογή στα στοιχεία της ευθείας παλινδρόμησης.



Διάγραμμα 10.3 Απεικόνιση της ευθείας ελαχίστων τετραγώνων

Σύμφωνα με τα παραπάνω, ισχύει η ακόλουθη σχέση:

$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$	**(10.32)**

Η παραπάνω σχέση ισχύει για κάθε σημείο (X_i, Y_i), οπότε ισχύει και για το άθροισμα των τετραγώνων των αντίστοιχων αποκλίσεων. Δηλαδή:

$\sum_i^n (Y_i - \bar{Y})^2 = \sum_i^n (\hat{Y}_i - \bar{Y})^2 + \sum_i^n (Y_i - \hat{Y}_i)^2$ ή SST=SSR+SSE	**(10.33)**

όπου:

- SST = $\sum_i^n (Y_i - \bar{Y})^2$ = Ολικό Άθροισμα των Τετραγώνων (Sum of Squares Total) ή συνολική διασπορά των δειγματικών τιμών της Y γύρω από τη μέση τιμή τους $\bar{Y}$, ή συνολική μεταβλητότητα εξαρτημένης μεταβλητής.

- SSR = $\sum_i^n (Y_i - \hat{Y}_i)^2$ = Άθροισμα Τετραγώνων της Παλινδρόμησης (Regression Sum of Squares) ή διασπορά των εκτιμημένων τιμών της Y γύρω από τη μέση τιμή $\bar{Y}$ που ερμηνεύεται από το υπόδειγμα παλινδρόμηση που εφαρμόσαμε ή μεταβλητότητα

εξαρτημένης μεταβλητής εξαιτίας της ανεξάρτητης μεταβλητής.

- SSE = $\sum_{i}^{n}\left(Y_i-\hat{Y}_i\right)^2$ = Άθροισμα Τετραγώνων των Σφαλμάτων (Error Sum of Squares) ή διασπορά των δειγματικών τιμών της Υ γύρω από την εκτιμημένη ευθεία παλινδρόμησης, δηλαδή ή ποσότητα που δεν ερμηνεύεται από το υπόδειγμα παλινδρόμηση που εφαρμόσαμε ή μεταβλητότητα της εξαρτημένης μεταβλητής εξαιτίας τυχαίων παραγόντων.

Ένας πίνακας ανάλυσης διακύμανσης της γραμμής παλινδρόμησης που περιλαμβάνει τα παραπάνω μεγέθη, έχει την ακόλουθη μορφή:

Πίνακας 10.3 Ανάλυσης διακύμανσης

Πηγή μεταβλητότητας (source of variation)	Άθροισμα τετραγώνων αποκλίσεων (Sum of Squares)	Βαθμοί ελευθερίας (Degrees of Freedom) d.f	Μέσο τετραγωνικό σφάλμα (mean square)	Στατιστική F
Παλινδρόμηση (Regression)	$SSR=\sum_{i}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	1	SSR/1	
Σφάλμα (Error)	$SSE=\sum_{i}^{n}\left(Y_i-\hat{Y}_i\right)^2$	n-2	SSE/(n-2)	$F=\dfrac{SSR/1}{SSE/(n-2)}$
Ολική (Total)	$SST=\sum_{i}^{n}\left(Y_i-\bar{Y}\right)^2$	n-1	SST/(n-1)	

Με βάση τα στοιχεία του πίνακα ανάλυσης διακύμανσης της γραμμής παλινδρόμησης και ειδικότερα με τη στατιστική F, μπορούμε να ελέγξουμε αν υπάρχει ή όχι γραμμική σχέση μεταξύ των μεταβλητών Υ και Χ. Για το σκοπό αυτό, ελέγχουμε την ακόλουθη στατιστική υπόθεση:

H_0: Μη γραμμική σχέση μεταξύ Υ και Χ

H_ε: Γραμμική σχέση μεταξύ Υ και Χ

Αν η τιμή της στατιστικής F είναι μεγαλύτερη από την τιμή $F_\alpha(1,n-2)$ η οποία προσδιορίζεται από τους στατιστικούς πίνακες της F κατανομής με 1 και n-2 βαθμούς ελευθερίας και επίπεδο σημαντικότητας α, τότε δεχόμαστε την εναλλακτική υπόθεση, δηλαδή ότι υπάρχει γραμμική σχέση μεταξύ των μεταβλητών Υ και Χ.

Παράδειγμα 10.7

Με τα δεδομένα του παραδείγματος 10.1, να κατασκευαστεί πίνακας ανάλυσης διακύμανσης και να ελεγχθεί αν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών και με πιθανότητα 95%.

Απάντηση:

Με βάση τα δεδομένα του προβλήματος, υπολογίζουμε το άθροισμα τετραγώνων των αποκλίσεων που ερμηνεύεται από το υπόδειγμα της παλινδρόμησης, το άθροισμα τετραγώνων των αποκλίσεων που δεν ερμηνεύεται από αυτό ως και τη συνολική διασπορά των τιμών της Y γύρω από το μέσο της:

Πίνακας 10.4 Εκτίμηση αθροίσματος τετραγώνων των αποκλίσεων (SSR, SSE, SST)

Πωλήσεις σε χιλ. € (Y)	Εισόδημα νοικοκυρών σε χιλ € (X)	$\hat{Y}_i$	$\sum_i^n (\hat{Y}_i - \bar{Y})^2$	$\sum_i^n (Y_i - \hat{Y}_i)^2$	$\sum_i^n (Y_i - \bar{Y})^2$
323.8	13.2	300.9	4.926.3	523.2	2.238.6
356.6	22.6	389.6	343.5	1.092.1	210.6
295,7	16.9	335.8	1.243.7	1.611.9	5.687.2
366.0	21.0	374.5	11.8	73.0	26.1
448.5	21.6	380,2	82,7	4.663,7	5.988.6
273.8	15,0	317,9	2.830,1	1.946,1	9.469.9
370,7	21.9	383.0	142.2	152.3	0.2
482.8	26.4	425.5	2.959.3	3.281.8	12.473.8
345.8	25,5	417,0	2.107.3	5.072.1	640.8
324.2	15.0	317.9	2.830.1	39.5	2.200.9
434.4	28.1	441.6	4.962.5	51.2	4.005.2
473,6	31.6	474.6	10.708.0	1.0	10.503.5
316,7	19.0	355.7	238.5	1.518.6	2.960.8
307,7	20.1	366.1	25.6	3.404.9	4.021.3
310.0	16.2	329.2	1.753.3	370.2	3.734.9
381.1	21.4	378.3	51.9	7.7	99.7
420,0	18,8	353,8	300,4	4.384,9	2.389,9
376.0	15.3	320.7	2.536.8	3.052.9	23.9
365.6	16.7	334.0	1.380.4	1.001.1	30.4
403.8	22.4	387.8	277.1	257.3	1.068.4
458.8	26.5	426.5	3.062.9	1.046.1	7.688.9
328,9	18,8	353,8	300,4	619,1	1.782,0
8.164,5	454,0	8.164,5	43.075,0	34.170,7	77.245,7
$\bar{Y} = 371{,}1$	$\bar{X} = 20{,}6$				

Επομένως, ο πίνακας ανάλυσης διακύμανσης των πωλήσεων της επιχείρησης, έχει ως εξής:

Πίνακας 10.5 Πίνακας ανάλυσης διακύμανσης

Πηγή μεταβλητότητας (source of variation)	Άθροισμα τετραγώνων αποκλίσεων (Sum of Squares)	Βαθμοί ελευθερίας (Degrees of Freedom) d.f	Μέσο τετραγωνικό σφάλμα (mean square)	Στατιστική F
Παλινδρόμηση (Regression)	$SSR=\sum_i^n(\hat{Y}_i-\bar{Y})^2=$ $=43.075,0$	1	SSR/1 = 43.075,0	$F=\dfrac{SSR/1}{SSE/(n-2)}=\dfrac{43.075,0}{1.708,54}=$ $=25,21$
Σφάλμα (Error)	$SSE=\sum_i^n(Y_i-\hat{Y}_i)^2=$ $=34.170,7$	n-2 = 22 – 2 = 20	SSE/(n-2) = 34.170,7/20 = = 1.708,54	
Ολική (Total)	$SST=\sum_i^n(Y_i-\bar{Y})^2=$ $=77.245,7$	n-1 = 22 – 1 = 21	SST/(n-1) = 77.245,7/21 = = 3.678,37	

Ο έλεγχος για την ύπαρξη ή μη γραμμικής σχέσης μεταξύ των μεταβλητών Y και X, θα γίνει με τη στατιστική F. Στη περίπτωση αυτή, οι υποθέσεις διατυπώνονται όπως παραπάνω. Επομένως, για επίπεδο σημαντικότητας α=0,05 και με βαθμούς ελευθερίας $v_1=1$ και $v_2=n-2=20$, υπολογίζεται από τους στατιστικούς πίνακες ότι $F_{0,05}(1,20)=4,35$.

Επειδή $F=25,21>F_{0,05}(1,20)=4,35$, απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική, που σημαίνει ότι υπάρχει γραμμική σχέση μεταξύ του εισοδήματος των νοικοκυριών (X) και των πωλήσεων της επιχείρησης (Y).

10.2.9 Συντελεστής προσδιορισμού (R)

Αξιοποιώντας τα στοιχεία που μας παρέχει ένας πίνακας ανάλυσης διακύμανσης, προκύπτει ο συντελεστής προσδιορισμού (coefficient of determination), ο οποίος συμβολίζεται με R^2 και ορίζεται από τη σχέση:

$$R^2=\frac{SSR}{SST}=\frac{SST\text{-}SSE}{SST}=1-\frac{SSE}{SST} \quad \text{ή} \quad R^2=1-\frac{\sum_i^n(Y_i-\hat{Y}_i)^2}{\sum_i^n(Y_i-\bar{Y})^2} \qquad \textbf{(10.34)}$$

Ο συντελεστής προσδιορισμού παίρνει τιμές από 0 έως 1, δηλαδή $0 \le R^2 \le 1$ και μετράει την αναλογία της συνολικής μεταβλητότητας γύρω από τη μέση τιμή $\bar{Y}$ που ερμηνεύεται από την παλινδρόμηση. Όσο πιο κοντά στη μονάδα είναι η τιμή του συντελεστή R^2,τόσο μεγαλύτερη είναι η ερμηνευτική ικανότητα του υποδείγματος παλινδρόμησης, δηλαδή η παλινδρόμηση εξηγεί μεγάλο ποσοστό της συνολικής διακύμανσης των παρατηρούμενων τιμών της Y. Αν $R^2=1$, τότε λέμε ότι υπάρχει τέλεια προσαρμογή, δηλαδή όλα τα σημεία των δεδομένων μας βρίσκονται πάνω στην ευθεία ελαχίστων τετραγώνων.

Εν αντιθέσει με τον συντελεστή προσδιορισμού R^2 που αναφέρεται στο δείγμα, αν διορθώσουμε το R^2 ως προς τους βαθμούς ελευθερίας του, τότε προκύπτει ο διορθωμένος συντελεστής προσδιορισμού $\bar{R}^2$ ($R^2 - adjusted$) ο οποίος ορίζεται ως:

$$\bar{R}^2 = \frac{SSR/(n-2)}{SST/(n-1)} = 1-(1-R^2)\frac{n-1}{n-2}$$	**(10.35)**

Παράδειγμα 10.8

Με τα δεδομένα του παραδείγματος 10.1 και δεδομένου ότι από το προηγούμενο παράδειγμα υπάρχει γραμμική σχέση μεταξύ των μεταβλητών εισόδημα νοικοκυριών και πωλήσεις, να εξεταστεί ο βαθμός της εξάρτησης των μεταβλητών.

Απάντηση:

Ο ζητούμενος βαθμός εξάρτησης των μεταβλητών εισόδημα νοικοκυριών και πωλήσεις, μπορεί να υπολογιστεί από τον συντελεστή προσδιορισμού R^2. Με βάση τα αποτελέσματα του πίνακα ανάλυσης της διακύμανσης (Πίνακας 10.4) προκύπτει ότι:

$$R^2 = 1-\frac{SSE}{SST} = 1-\frac{34.170,7}{77.245,7} = 1-0,44 = 0,56$$

Αυτό σημαίνει ότι, οι πωλήσεις τις επιχείρησης εξαρτώνται από το εισόδημα των νοικοκυριών κατά 56%, ενώ όλοι οι άλλοι παράγοντες που δεν λάβαμε υπόψη μας στο συγκεκριμένο υπόδειγμα παλινδρόμησης επιδρούν κατά 44%. Με άλλα λόγια, το 56% της μεταβλητικότητας των πωλήσεων οφείλεται στην επίδραση του εισοδήματος των νοικοκυριών.

10.3 Συσχέτιση

Στο προηγούμενο κεφάλαιο, ασχοληθήκαμε με την ανάλυση παλινδρόμησης και συγκεκριμένα με τη μέθοδο των ελαχίστων τετραγώνων και εκτιμήσαμε τη σχέση μιας μεταβλητής ως προς μία άλλη, υποθέτοντας ότι μεταξύ αυτών των δύο μεταβλητών υπάρχει γραμμική σχέση. Επιπλέον, βασικό χαρακτηριστικό της ανάλυσης παλινδρόμησης ήταν ο χαρακτηρισμός των μεταβλητών Y και X, ως εξαρτημένη τυχαία μεταβλητή η πρώτη και ανεξάρτητη μη τυχαία μεταβλητή η δεύτερη.

Υπάρχουν όμως περιπτώσεις που επιθυμούμε να διερευνήσουμε τη σχέση μεταξύ δύο μεταβλητών, αδυνατώντας να θεωρήσουμε τη μια μεταβλητή ως εξαρτημένη και την άλλη ως ανεξάρτητη, αφού καμία δεν πληροί τις σχετικές προϋποθέσεις. Στην περίπτωση αυτή, με τη μέθοδο της συσχέτισης, μπορούμε να μετρήσουμε τον βαθμό ή την ένταση της υφιστάμενης γραμμικής σχέσης μεταξύ των δύο (ή περισσοτέρων) μεταβλητών, αρκεί όλες να είναι στοχαστικές ή τυχαίες μεταβλητές. Η μέτρηση της σχέσης αυτής, γίνεται με τον λεγόμενο συντελεστή συσχέτισης ο οποίος είναι ανεξάρτητος των μονάδων μέτρησης των μεταβλητών, δηλαδή είναι ένας καθαρός αριθμός επιτρέποντας έτσι τις συγκρίσεις. Επιπλέον, ο συντελεστής συσχέτισης συνδέεται άμεσα με την εξίσωση παλινδρόμησης, αφού εξαρτάται από το τυπικό σφάλμα εκτίμησης και αποτελεί ένα μέτρο καλής προσαρμογής της στις παρατηρήσεις του δείγματος.

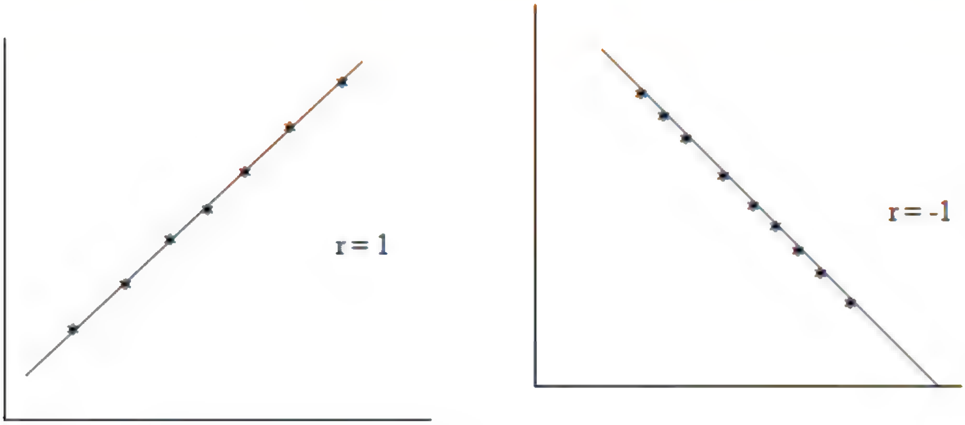
10.3.1 Συντελεστής γραμμικής συσχέτισης

Ο συντελεστής συσχέτισης δύο τυχαίων μεταβλητών, είναι ένα μέτρο που μετρά τον βαθμό της συμμεταβολής των δύο μεταβλητών ή την ένταση της γραμμικής σχέσης που πιθανόν να υπάρχει μεταξύ τους. Για το λόγο αυτό, ονομάζεται και συντελεστής γραμμικής συσχέτισης του Pearson ή δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson, συμβολίζεται με το γράμμα r και ορίζεται από τον τύπο:

$$r = \frac{\sum_{i}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i}^{n}(X_i - \bar{X})^2 \sum_{i}^{n}(Y_i - \bar{Y})^2}} = \frac{n\sum_{i}^{n}X_iY_i - \sum_{i}^{n}X_i\sum_{i}^{n}Y_i}{\sqrt{\left[n\sum_{i}^{n}X_i^2 - \left(\sum_{i}^{n}X_i\right)^2\right]\left[n\sum_{i}^{n}Y_i^2 - \left(\sum_{i}^{n}Y_i\right)^2\right]}} \qquad (10.36)$$

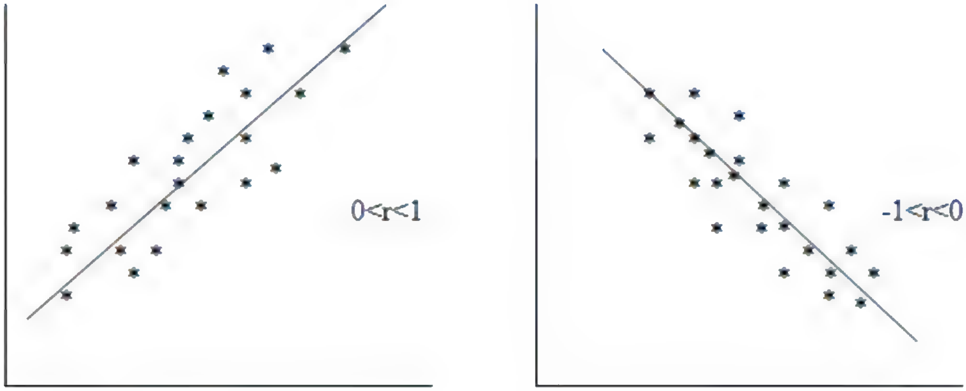
Ο πληθυσμιακός συντελεστής γραμμικής συσχέτισης του Pearson συμβολίζεται με το γράμμα ρ και ορίζεται ανάλογα, αλλά ο υπολογισμός του είναι δύσκολος γιατί απαιτεί τη χρησιμοποίηση ολόκληρου του πληθυσμού, πράγμα ασύμφορο γι' αυτό και χρησιμοποιούμε τον r. Κατά συνέπεια ο r είναι εκτιμητής του ρ και εξαρτάται από το μέγεθος του δείγματος και την κοινή κατανομή των μεταβλητών X και Y, που σημαίνει ότι όταν η πληθυσμιακή κατανομή αυτών είναι κανονική τότε ο r θεωρείται αξιόπιστος.

Ο συντελεστής συσχέτισης του Pearson παίρνει τιμές από -1 έως +1, δηλαδή στο διάστημα $-1 \le r \le +1$. Για το διάστημα αυτό των τιμών του r μπορούμε να παρατηρήσουμε τα εξής:
α) Αν r=+1 και r=-1, τότε λέμε ότι έχουμε αντίστοιχα τέλεια θετική και αρνητική γραμμική συσχέτιση. Στην περίπτωση αυτή, τα σημεία του διαγράμματος διασποράς βρίσκονται όλα πάνω σε μια ευθεία γραμμή με θετική και αρνητική κλίση αντίστοιχα.



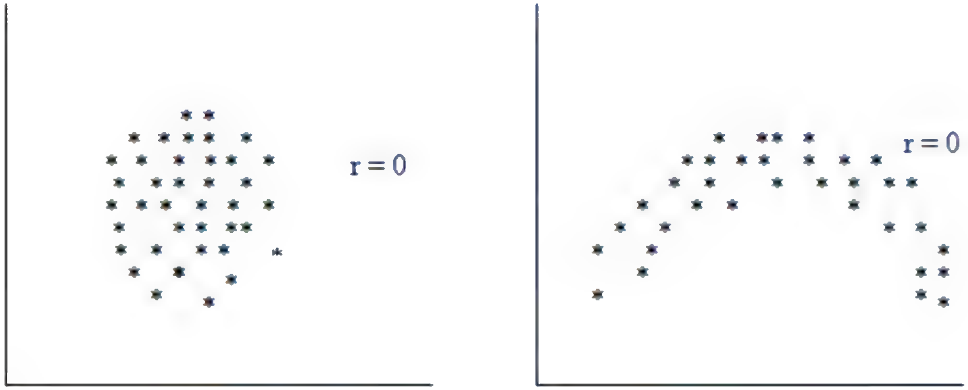
Διάγραμμα 10.4 Διάγραμμα Διασποράς

β) Αν 0<r<+1 και -1<r<0, τότε λέμε ότι έχουμε αντίστοιχα θετική γραμμική και αρνητική γραμμική συσχέτιση. Αν σε περίπτωση η τιμή του r πλησιάζει το +1 ή το -1, τότε λέμε ότι έχουμε ισχυρή θετική και ισχυρή αρνητική συσχέτιση αντίστοιχα. Αντιθέτως, αν η τιμή του r πλησιάζει το 0, τότε λέμε ότι έχουμε ασθενή θετική και ασθενή αρνητική συσχέτιση αντίστοιχα.



Διάγραμμα 10.5 Διάγραμμα Διασποράς

γ) Αν r=0, τότε λέμε ότι έχουμε μηδενική γραμμική συσχέτιση, δηλαδή δεν υπάρχει γραμμική συσχέτιση. Στην περίπτωση αυτή, τα σημεία του διαγράμματος διασποράς είτε είναι σκορπισμένα στο χώρο, είτε δημιουργούν εικόνα τετραγωνικής σχέσης μεταξύ των δύο μεταβλητών (με τη γραμμική συσχέτιση να είναι μηδέν).



Διάγραμμα 10.5 Διάγραμμα Διασποράς

Παράδειγμα 10.9

Με τα δεδομένα του παραδείγματος 10.1., να εκτιμηθεί ο βαθμός συσχέτισης ή συμμεταβολής των μεταβλητών εισόδημα νοικοκυριών και πωλήσεις της επιχείρησης.

Απάντηση:

Ο ζητούμενος βαθμός συσχέτισης των μεταβλητών εισόδημα νοικοκυριών και πωλήσεις, μπορεί να υπολογιστεί από τον συντελεστή γραμμικής συσχέτισης του Pearson. Με βάση τα στοιχεία του Πίνακα 1, προκύπτει ότι:

$$r = \frac{n\sum_{i}^{n} X_i Y_i - \sum_{i}^{n} X_i \sum_{i}^{n} Y_i}{\sqrt{\left[n\sum_{i}^{n} X_i^2 - \left(\sum_{i}^{n} X_i\right)^2\right]\left[n\sum_{i}^{n} Y_i^2 - \left(\sum_{i}^{n} Y_i\right)^2\right]}} =$$

$$= \frac{22 \cdot 173.049,4 - 454,0 \cdot 8.164,5}{\sqrt{\left[22 \cdot 9.852,4 - (454,0)^2\right]\left[22 \cdot 3.107.203,0 - (8.164,5)^2\right]}} = 0,75$$

Επομένως, υπάρχει μια σημαντική ή ισχυρή θετική συσχέτιση μεταξύ των μεταβλητών Χ και Υ, που σημαίνει ότι νοικοκυριά με υψηλό μέσο οικογενειακό εισόδημα καταναλώνουν περισσότερο οπότε και αυξάνουν οι πωλήσεις τις επιχείρησης, ενώ αντιθέτως νοικοκυριά με χαμηλότερο μέσο οικογενειακό εισόδημα καταναλώνουν λιγότερο οι πωλήσεις τις επιχείρησης παρουσιάζονται περιορισμένες.

Στο σημείο αυτό, αξίζει ν' αναφέρουμε ότι το τετράγωνο του συντελεστή γραμμικής συσχέτισης του Pearson, είναι ο συντελεστής προσδιορισμού R^2 που αναφέραμε παραπάνω. Δηλαδή,

$$r^2 = 0,75^2 = 0,56 = R^2$$

10.3.2 Συσχέτιση και παλινδρόμηση

Ο δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson, μπορεί να οριστεί και από τον τύπο:

$r = \dfrac{s_{XY}}{s_X \cdot s_Y}$	**(10.37)**

όπου,

$s_{XY} = Cov(X,Y) = \dfrac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \dfrac{\sum_i^n X_i Y_i - n\bar{X}\bar{Y}}{n-1}$	**(10.38)**
$s_X = \sqrt{\dfrac{1}{n-1}\sum_i^n (X_i - \bar{X})^2}$ (τυπική απόκλιση του X)	**(10.39)**
$s_Y = \sqrt{\dfrac{1}{n-1}\sum_i^n (Y_i - \bar{Y})^2}$ (τυπική απόκλιση του Y)	**(10.40)**

Σύμφωνα με τα παραπάνω εκτεθέντα, ο συντελεστής παλινδρόμησης $\hat{\beta}$ (υποθέτοντας ότι εκτελούμε παλινδρόμηση της Y πάνω στη X) δίνεται από τη σχέση:

$\hat{\beta} = \dfrac{s_{XY}}{s_X^2} = \dfrac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i^n (X_i - \bar{X})^2} = \dfrac{\sum_i^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_i^n X_i^2 - n\bar{X}^2}$ ή $\hat{\beta} = \dfrac{r \cdot s_Y}{s_X}$	**(10.41)**

Η σχέση αυτή ορίζεται ανάλογα αν υποθέσουμε ότι εκτελούμε παλινδρόμηση της Χ πάνω στη Υ.

Τέλος, αποδεικνύεται ότι ο συντελεστής προσδιορισμού της γραμμής παλινδρόμησης δύο μεταβλητών ισούται με το τετράγωνο του συντελεστή γραμμικής συσχέτισης του Pearson. Δηλαδή:

$R^2 = r^2$	**(10.42)**

10.3.3 Έλεγχος υποθέσεων συντελεστή γραμμικής συσχέτισης

Όπως αναφέραμε παραπάνω, ο συντελεστής γραμμικής συσχέτισης r αποτελεί εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης ρ και επομένως θα πρέπει να αξιολογηθεί για να διαπιστώσουμε αν η εκτίμηση είναι καλή ή όχι. Με άλλα λόγια, θα πρέπει να ελέγξουμε με κάποια πιθανότητα αν ο συντελεστής συσχέτισης που εκτιμήσαμε είναι στατιστικά σημαντικός και κατά συνέπεια αξιόπιστος.

Στην περίπτωση αυτή, για να εκτελέσουμε ελέγχους υποθέσεων θα πρέπει να γνωρίζουμε την κατανομή δειγματοληψίας του r. Αν όμως θεωρήσουμε ότι η από κοινού κατανομή των (Χ,Υ) ακολουθεί την κανονική κατανομή, τότε γνωρίζουμε την κατανομή δειγματοληψίας του r. Όμως, επειδή η μορφή της κατανομής του r διαφοροποιείται ανάλογα με το αν ο πληθυσμιακός συντελεστής είναι ίσος με το μηδέν ή διάφορος του μηδενός (ρ=0 ή ρ≠0), τότε διακρίνουμε τις ακόλουθες περιπτώσεις ελέγχων:

1. Για ρ=0

Στην περίπτωση αυτή, αν η από κοινού κατανομή των (Χ,Υ) ακολουθεί την κανονική κατανομή και επιπλέον η μια είναι ανεξάρτητη της άλλης (δηλαδή ρ=0), τότε η κατανομή του r είναι συμμετρική με μέσο μηδέν και διακύμανση $\frac{1-r^2}{n-2}$ η οποία εξαρτάται από το μέγεθος του δείγματος n. Η διατύπωση των υποθέσεων, εξειδικεύεται ως ακολούθως:

Διατύπωση υποθέσεων

H_0: Οι μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες
(δεν υπάρχει συσχέτιση μεταξύ τους)
Δηλαδή ρ=0

H_ε: i) Είτε υπάρχει τάση οι μεγαλύτερες τιμές της X να αντιστοιχούν στις μεγαλύτερες της Y,
Είτε υπάρχει τάση στις μικρότερες τιμές της X να αντιστοιχούν οι μεγαλύτερες της Y
Δηλαδή $\rho \neq 0$

ii) Είτε υπάρχει τάση οι μεγαλύτερες τιμές της X να αντιστοιχούν στις μεγαλύτερες της Y (θετική συσχέτιση),
Είτε υπάρχει τάση στις μικρότερες τιμές της X να αντιστοιχούν οι μικρότερες της Y (θετική συσχέτιση)
Δηλαδή ρ>0

iii) Είτε υπάρχει τάση οι μικρότερες τιμές της X να αντιστοιχούν στις μεγαλύτερες της Y (αρνητική συσχέτιση),
Είτε υπάρχει τάση στις μεγαλύτερες τιμές της X να αντιστοιχούν οι μικρότερες της Y (αρνητική συσχέτιση)
Δηλαδή ρ<0

Στατιστική συνάρτηση ελέγχου

Αν το μέγεθος του δείγματος είναι σχετικά μικρό, τότε η τυχαία μεταβλητή

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \textbf{(10.43)}$$

ακολουθεί την κατανομή t-student με ν=n-2 βαθμούς ελευθερίας. Η στατιστική αυτή, μπορεί να χρησιμοποιηθεί ακόμα και στην περίπτωση που αδυνατούμε να υποθέσουμε ότι η κοινή κατανομή των X,Y είναι κανονική, αρκεί να ισχύει ότι n≥30.

Απόφαση

Απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική Η σε επίπεδο σημαντικότητας α, αν για τα παραπάνω τρία είδη ελέγχου ισχύει αντίστοιχα:

i) $t < -t_{\nu,\alpha/2}$ ή $t > t_{\nu,\alpha/2}$, ii) $t \geq t_{\nu,\alpha/2}$ και iii) $t \leq -t_{\nu,\alpha/2}$.

2. Για $\rho=\rho_0$

Για να ελέγξουμε την υπόθεση ότι ο πληθυσμιακός συντελεστής συσχέτισης ρ έχει δεδομένη τιμή διάφορη του μηδενός, δεν μπορούμε να χρησιμοποιήσουμε την κατανομή δειγματοληψίας του r και κατά συνέπεια τη στατιστική t-student (όπως ανωτέρω), διότι η μορφή της κατανομής του r είναι ασύμμετρη. Στην περίπτωση αυτή, χρησιμοποιούμε το προσεγγιστικό κριτήριο του Fisher.

Διατύπωση υποθέσεων

$H_0: \rho=\rho_0$

H_z : i) $\rho \neq \rho_0$

ii) $\rho>\rho_0$

iii) $\rho<\rho_0$

Στατιστική συνάρτηση ελέγχου

Σύμφωνα με τον Fisher, ο μετασχηματισμός $z_r = \frac{1}{2}\ln\frac{1+r}{1-r}$ ακολουθεί προσεγγιστικά την κανονική κατανομή με μέσο $\mu_z = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}$ και διακύμανση $\sigma_z^2 = \frac{1}{n-3}$ η οποία είναι ανεξάρτητη της τιμής του r για οποιαδήποτε τιμή του ρ_0 και n. Επομένως, η στατιστική ελέγχου που χρησιμοποιείται είναι:

$z = \frac{z_r - z_\rho}{\sigma_z} = \frac{\left(\frac{1}{2}\ln\frac{1+r}{1-r}\right) - \left(\frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}} \sim N(0,1)$	**(10.44)**

Απόφαση

Απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική H_z σε επίπεδο σημαντικότητας α, αν για τα παραπάνω τρία είδη ελέγχου ισχύει αντίστοιχα:

i) $z \leq -z_{\alpha/2}$ ή $z \geq z_{\alpha/2}$, ii) $z \geq z_{\alpha/2}$ και iii) $z \leq -z_{\alpha/2}$

Παράδειγμα 10.10

Να ελεγχθεί σε επίπεδο σημαντικότητας α = 0,05, αν οι μεταβλητές μέσο οικογενειακό εισόδημα νοικοκυριών και πωλήσεις της επιχείρησης, συσχετίζονται στον πληθυσμό απ' όπου προήλθε το δείγμα.

Απάντηση:

Στην περίπτωση αυτή, ο έλεγχος διατυπώνεται ως εξής

$H_0: \rho = 0$

$H_1: \rho \neq 0$

Η μηδενική υπόθεση μας λέει ότι ο συντελεστής συσχέτισης του πληθυσμού είναι μηδέν ή με άλλα λόγια ότι οι μεταβλητές μέσο οικογενειακό εισόδημα νοικοκυριών και πωλήσεις είναι αμοιβαία ανεξάρτητες, δηλαδή δεν υπάρχει συσχέτιση μεταξύ τους.

Από το παράδειγμα 10.9, προέκυψε ότι ο συντελεστής συσχέτισης για τα δεδομένα του δείγματος είναι r=0,75. Για τον έλεγχο που διατυπώσαμε παραπάνω, θα χρησιμοποιήσουμε τη στατιστική t-student. Δηλαδή:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,75\sqrt{20}}{\sqrt{1-0,56}} = 5,08 .$$

Από τον Πίνακα κριτικών τιμών της κατανομής t-student και για α/2=0,025 και ν = 22-2 = 20, βρίσκουμε ότι $t_{20,0.025} = 2,086$.

Επομένως, αφού $t = 5,08 > t_{20,0.025} = 2,086$, αυτό σημαίνει ότι απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική, δηλαδή συμπεραίνουμε ότι το μέσο οικογενειακό εισόδημα νοικοκυριών και οι πωλήσεις συσχετίζονται μεταξύ τους στον πληθυσμό απ' όπου προήλθε το δείγμα των παρατηρήσεων.

10.4 Πολλαπλή γραμμική παλινδρόμηση

10.4.1 Υπόδειγμα πολλαπλής γραμμικής παλινδρόμησης

Στην προηγούμενη ενότητα περιγράψαμε τη μέθοδο της απλής γραμμικής παλινδρόμησης και συσχέτισης η οποία περιορίστηκε στην ανάλυση της σχέσης μεταξύ δύο μεταβλητών. Ειδικότερα, έγινε η παραδοχή ότι η μια μεταβλητή εξαρτάται μόνο από μια άλλη, αγνοώντας τυχόν επιδράσεις άλλων μεταβλητών στην εξαρτημένη μεταβλητή.

Πολλές φορές όμως, η λύση επιχειρηματικών, κοινωνικών κ.λπ. προβλημάτων βασίζεται στη σχέση μεταξύ τριών, τεσσάρων ή και περισσοτέρων μεταβλητών. Για παράδειγμα, η αμοιβή ενός εργαζομένου δεν επηρεάζεται μόνο από την ηλικία του, αλλά και από μια σειρά άλλων παραγόντων όπως, το επίπεδο εκπαίδευσης, το φύλο (άνδρας ή γυναίκα), τα χρόνια εμπειρίας στην παρούσα ή σε προηγούμενη εργασία, την οικογενειακή του

κατάσταση (παντρεμένος, ανύπαντρος, διαζευγμένος) κ.λπ. Για την επίλυση λοιπόν ενός προβλήματος προσδιορισμού των αμοιβών από μια σειρά παραγόντων, θα πρέπει να επεκτείνουμε το υπόδειγμα απλής παλινδρόμησης ώστε να περιλαμβάνει και τις υπόλοιπες ερμηνευτικές μεταβλητές.

Η αναγκαιότητα επέκτασης της ανάλυσης απλής παλινδρόμησης και συσχέτισης σε υποδείγματα στα οποία μια μεταβλητή εξαρτάται (επεξηγείται) από περισσότερες από μια ανεξάρτητες μεταβλητές, παραπέμπει στην ανάλυση πολλαπλής παλινδρόμησης και πολλαπλής συσχέτισης.

Ένα υπόδειγμα πολλαπλής παλινδρόμησης που περιγράφει τη σχέση μεταξύ της εξαρτημένης μεταβλητής Y_i και των k ανεξάρτητων μεταβλητών $X_1,X_2,\ldots,X_k$ έχει την εξής μορφή:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \qquad \textbf{(10.45)}$$

όπου:
Y_i = η τιμή της εξαρτημένης μεταβλητής.

$X_{1i},X_{2i},\ldots,X_{ki}$= οι παρατηρήσεις των ανεξάρτητων μεταβλητών $X_1,X_2,\ldots,X_k$ αντίστοιχα. Ειδικότερα, ο πρώτος δείκτης αναφέρεται στην ερμηνευτική μεταβλητή και ο δεύτερος στην παρατήρηση.

β_0 = μία σταθερά.

$\beta_1,\beta_2,\ldots,\beta_k$ = οι πληθυσμιακοί συντελεστές παλινδρόμησης που περιγράφουν την επίδραση των ανεξάρτητων μεταβλητών στην εξαρτημένη.

ε_i = μια τυχαία μεταβλητή που ονομάζεται σφάλμα ή κατάλοιπο ή διαταρακτικός όρος.

Το παραπάνω υπόδειγμα πολλαπλής παλινδρόμησης, βασίζεται στις ακόλουθες υποθέσεις:

1. Οι ανεξάρτητες μεταβλητές είναι γνωστές σταθερές και μπορούν να μετρηθούν χωρίς κάποιο σφάλμα.
2. Οι ανεξάρτητες είναι γραμμικά ανεξάρτητες.
3. Οι τιμές της εξαρτημένης μεταβλητής είναι όλες ανεξάρτητες μεταξύ τους.
4. Οι διακυμάνσεις των κατανομών της τυχαίας μεταβλητής είναι όλες ίσες μεταξύ τους.
5. Σε κάθε σύνολο τιμών των ανεξάρτητων μεταβλητών (i=1,2,...,k) αντιστοιχεί μια κατανομή της τυχαίας μεταβλητής.
6. Τα σφάλματα είναι ανεξάρτητα μεταξύ τους και κατανέμονται κανονικά.
7. Οι μέσες τιμές (αναμενόμενες τιμές) των σφαλμάτων είναι μηδέν.
8. Τα σφάλματα έχουν την ίδια διακύμανση για όλους τους συνδυασμούς των τιμών των ανεξάρτητων μεταβλητών.

10.4.2 Εκτίμηση της εξίσωσης της πολλαπλής γραμμικής παλινδρόμησης

Όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης έτσι και εδώ, σκοπός είναι να εκτιμήσουμε τις πληθυσμιακές παραμέτρους (συντελεστές) $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ της πολλαπλής παλινδρόμησης από τα δεδομένα του δείγματος. Αν συμβολίσουμε με $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ τους συντελεστές πολλαπλής παλινδρόμησης του δείγματος, τότε η εξίσωση που θα προκύψει από την εκτίμηση των συντελεστών αυτών γράφεται ως εξής:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$	**(10.46)**

Σ' ένα υπόδειγμα πολλαπλής γραμμικής παλινδρόμησης, οι συντελεστές $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ ονομάζονται συντελεστές μερικής παλινδρόμησης διότι δείχνουν τη μερική επίδραση που ασκούν οι ανεξάρτητες μεταβλητές στην εξαρτημένη μεταβλητή.

Η εκτίμηση της εξίσωσης (10.46) προκύπτει με τη χρησιμοποίηση της γνωστής μεθόδου των ελαχίστων τετραγώνων σύμφωνα με την οποία αναζητούμε εκείνες τις τιμές των $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ που ελαχιστοποιούν το άθροισμα των κάθετων τετραγωνικών αποκλίσεων ανάμεσα στις πραγματικές τιμές του δείγματος της εξαρτημένης μεταβλητής Y_i και τις υπολογισθείσες (θεωρητικές) τιμές $\hat{Y}_i$ που προκύπτουν από την εξίσωση παλινδρόμησης.

Στη συνέχεια θα εξετάσουμε την περίπτωση ενός υποδείγματος με δύο ανεξάρτητες μεταβλητές X_1 και X_2. Επομένως, το υπόδειγμα που θα εκτιμήσουμε είναι:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$	**(10.47)**

Οι εκτιμητές $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ του υποδείγματος αυτού είναι:

$\hat{\beta}_1 = \dfrac{\sum x_2^2 \sum x_1 y - \sum x_1 x_2 \sum x_2 y}{\sum x_1^2 \sum x_2^2 - \left(\sum x_1 x_2\right)^2}$ $\hat{\beta}_2 = \dfrac{\sum x_1^2 \sum x_2 y - \sum x_1 x_2 \sum x_1 y}{\sum x_1^2 \sum x_2^2 - \left(\sum x_1 x_2\right)^2}$ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$	**(10.48)**

Όπου

$x_1 = X_1 - \bar{X}_1, \quad x_2 = X_2 - \bar{X}_2, \quad y = Y - \bar{Y}$

10.4.3 Τυπικό σφάλμα εκτίμησης

Όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης, έτσι και εδώ το τυπικό σφάλμα εκτίμησης είναι ένα μέτρο αξιολόγησης του εκτιμηθέντος υποδείγματος, ώστε να γνωρίζουμε κατά πόσο καλά αυτό παριστά τα δεδομένα μας ή κατά πόσο αυτό προσαρμόζει τη θεωρητική στην πραγματική κατάσταση.
Για να υπολογίσουμε το τυπικό σφάλμα εκτίμησης, θα πρέπει κατ' αρχάς να εκτιμήσουμε τη μέση απόκλιση τετραγώνου των τιμών Y_i (παρατηρήσεις της Y) από τις εκτιμηθέντες τιμές $\hat{Y}_i$, δηλαδή τη διακύμανση των καταλοίπων γύρω από τη γραμμή παλινδρόμησης:

$$s^2_{Y/X\,X} = \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-k-1} = \frac{\sum y^2 - \hat{\beta}_1 \sum x_1 y - \hat{\beta}_2 \sum x_2 y}{n-k-1} \qquad \textbf{(10.49)}$$

Όπου:

$x_1 = X_1 - \bar{X}_1, \quad x_2 = X_2 - \bar{X}_2, \quad y = Y - \bar{Y}$

Το $s^2_{Y/X\,X}$, δίδει μια εκτίμηση της διασποράς των τυχαίων σφαλμάτων και ονομάζεται μέσο τετραγωνικό σφάλμα (mean squared error). Η θετική τετραγωνική ρίζα της διασποράς των τυχαίων σφαλμάτων ($\sqrt{s^2_{Y/X\,X}}$), ονομάζεται τυπικό σφάλμα εκτίμησης (standard eror of estimate) ή τυπική απόκλιση των καταλοίπων (residual standard deviation).

10.4.4 Ανάλυση διακύμανσης

Όπως και στην περίπτωση του απλού γραμμικού υποδείγματος, έτσι και στο υπόδειγμα της πολλαπλής γραμμικής παλινδρόμησης ισχύει:

$$\sum_i^n (Y_i - \bar{Y})^2 = \sum_i^n (\hat{Y}_i - \bar{Y})^2 + \sum_i^n (Y_i - \hat{Y}_i)^2 \quad \text{ή} \quad SST=SSR+SSE \qquad \textbf{(10.50)}$$

Όπου (για k=2):

- SST = $SST = \sum_i^n (Y_i - \bar{Y})^2 = \sum_i^n Y_i^2 - \frac{1}{n}\left(\sum_i^n Y\right)^2 = \sum y^2$ = Ολικό Άθροισμα των Τετραγώνων (Sum of Squares Total).

- SSR = $SSR = \sum_{i}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \hat{\beta}_1 \sum x_1 y + \hat{\beta}_2 \sum x_2 y$ = Άθροισμα Τετραγώνων της Παλινδρόμησης (Regression Sum of Squares) ή διασπορά των εκτιμημένων τιμών της Y γύρω από τη μέση τιμή $\bar{Y}$ που ερμηνεύεται από το υπόδειγμα παλινδρόμησης που εφαρμόσαμε.
- SSE = $\sum_{i}^{n}\left(Y_i - \hat{Y}_i\right)^2 = SST - SSR = \sum y^2 - \hat{\beta}_1 \sum x_1 y - \hat{\beta}_2 \sum x_2 y$ = Άθροισμα Τετραγώνων των Σφαλμάτων (Error Sum of Squares) ή διασπορά των δειγματικών τιμών της γύρω από την εκτιμημένη ευθεία παλινδρόμησης ή ποσότητα που δεν ερμηνεύεται από το υπόδειγμα παλινδρόμησης που εφαρμόσαμε.

Ένας πίνακας ανάλυσης διακύμανσης της γραμμής παλινδρόμησης που περιλαμβάνει τα παραπάνω μεγέθη, έχει την ακόλουθη μορφή:

Πηγή μεταβλητότητας (source of variation)	Αθροισμα τετραγώνων αποκλίσεων (Sum of Squares)	Βαθμοί ελευθερίας (Degrees of Freedom) d.f	Μέσο τετραγωνικό σφάλμα (mean square)	Στατιστική F
Παλινδρόμηση (Regression)	$SSR = \sum_{i}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	$k = 2$	SSR/2	$F = \frac{SSR/2}{SSE/(n-3)}$
Σφάλμα (Error)	$SSE = \sum_{i}^{n}\left(Y_i - \hat{Y}_i\right)^2$	n-k-1 =n-3	SSE/(n-3)	
Ολική (Total)	$SST = \sum_{i}^{n}\left(Y_i - \bar{Y}\right)^2$	n-1	SST/(n-1)	

Με βάση τα στοιχεία του πίνακα ανάλυσης διακύμανσης της γραμμής παλινδρόμησης και ειδικότερα με τη στατιστική F, μπορούμε να ελέγξουμε αν το ποσοστό των μεταβολών της Y που οφείλεται στις επιδράσεις των ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_k$ και εξηγείται από την εξίσωση πολλαπλής παλινδρόμησης είναι διάφορο του μηδενός. Για το σκοπό αυτό, ελέγχουμε την ακόλουθη στατιστική υπόθεση:

H_0: $\beta_1 = \beta_2 = 0$ (Η εξίσωση παλινδρόμησης δεν εξηγεί καθόλου τις μεταβολές της Y)

H_ε: Τουλάχιστον ένας συντελεστής $\beta_i \neq 0$ με i=1,2

(Η εξίσωση παλινδρόμησης εξηγεί ένα μεγάλο μέρος των μεταβολών της Y)

Αν η τιμή της στατιστικής F είναι μεγαλύτερη από την τιμή $F_\alpha(k, n-k-1)$ η οποία προσδιορίζεται από τους στατιστικούς πίνακες της F κατανομής με k και n-k-1 βαθμούς ελευθερίας και επίπεδο σημαντικότητας α, τότε απορρίπτουμε την H_0 και δεχόμαστε την εναλλακτική υπόθεση, δηλαδή ότι η εξίσωση παλινδρόμησης εξηγεί ένα μεγάλο μέρος των μεταβολών της Y.

10.4.5 Συντελεστής πολλαπλού προσδιορισμού

Αξιοποιώντας τα στοιχεία που μας παρέχει ένας πίνακας ανάλυσης διακύμανσης, προκύπτει ο συντελεστής πολλαπλού προσδιορισμού (multiple determination coeffi-cient), ο οποίος συμβολίζεται με R^2 και ορίζεται από τη σχέση:

$R^2 = \frac{SSR}{SST} = \frac{SST\text{-}SSE}{SST} = 1 - \frac{SSE}{SST}$ ή $R^2 = 1 - \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{\sum_i^n (Y_i - \bar{Y})^2}$ ή $R^2 = \frac{\hat{\beta}_1 \sum x_1 y + \hat{\beta}_2 \sum x_2 y}{\sum y^2}$	**(10.51)**

Ο συντελεστής προσδιορισμού παίρνει τιμές από 0 έως 1, δηλαδή $0 \leq R^2 \leq 1$ και μετράει την αναλογία της συνολικής μεταβλητότητας γύρω από τη μέση τιμή $\bar{Y}$ που ερμηνεύεται από την παλινδρόμηση. Όσο πιο κοντά στη μονάδα είναι η τιμή του συντελεστή R^2, τόσο μεγαλύτερη είναι η ερμηνευτική ικανότητα του υποδείγματος πολλαπλής παλινδρόμησης, δηλαδή η παλινδρόμηση εξηγεί μεγάλο ποσοστό της συνολικής διακύμανσης των παρατηρούμενων τιμών της Y. Αν $R^2 = 1$, τότε λέμε ότι υπάρχει τέλεια προσαρμογή.

Εν αντιθέσει με τον συντελεστή πολλαπλού προσδιορισμού R^2 που αναφέρεται στο δείγμα, αν διορθώσουμε το R^2 ως προς τους βαθμούς ελευθερίας του, τότε προκύπτει ο διορθωμένος συντελεστής προσδιορισμού $\bar{R}^2$ (R^2-adjusted) ο οποίος ορίζεται ως:

$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$	**(10.52)**

Όταν ελέγχουμε την υπόθεση ότι η εξίσωση πολλαπλής παλινδρόμησης εξηγεί ή όχι μεγάλο μέρος των μεταβολών της Y (έλεγχος που αναφέραμε παραπάνω με τη στατιστική F), είναι σαν να ελέγχουμε για τον πληθυσμό την υπόθεση:

$$H_0\colon R^2 = 0$$

$$H_1\colon R^2 \neq 0$$

Αν η τιμή της στατιστικής F είναι μεγαλύτερη από την τιμή $F_\alpha(k, n-k-1)$ η οποία προσδιορίζεται από τους στατιστικούς πίνακες της F κατανομής με k και n-k-1 βαθμούς ελευθερίας και επίπεδο σημαντικότητας α, τότε απορρίπτουμε την H_0 και δεχόμαστε την εναλλακτική υπόθεση, δηλαδή η εξίσωση παλινδρόμησης εξηγεί ένα μεγάλο μέρος των μεταβολών της Y ή ότι η μεταβλητή Y συνδέεται γραμμικά με τις μεταβλητές $X_1, X_2, .., X_k$.

10.4.6 Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων των παραμέτρων

Για να ελέγξουμε τους συντελεστές της β_i για i=1,2,...,k (δεδομένων των υποθέσεων του υποδείγματος της πολλαπλής γραμμικής παλινδρόμησης), θα πρέπει να γνωρίζουμε την κοινή διακύμανση σ^2 των κατανομών της Y. Συνήθως, επειδή η διακύμανση αυτή είναι άγνωστη, χρησιμοποιούμε μια αμερόληπτη εκτίμησή της η οποία είναι:

$s^2_{Y/X\,X} = \frac{1}{n-k-1}\sum_i^n (Y_i - \hat{Y}_i)^2$	**(10.53)**

Στην περίπτωση που έχουμε μόνο δύο ανεξάρτητες μεταβλητές, τότε η αμερόληπτη εκτίμηση της διακύμανσης σ^2 δίνεται από τη σχέση:

$s^2_{Y/X\,X} = \frac{\sum_i^n (Y_i-\bar{Y}_i)^2 - \hat{\beta}_1\sum_i^n (X_1-\bar{X}_1)(Y_i-\bar{Y}) - \hat{\beta}_2\sum_i^n (X_2-\bar{X}_2)(Y_i-\bar{Y})}{n-3}$ ή $s^2_{Y/X\,X} = \frac{\sum_i^n y^2 - \hat{\beta}_1\sum_i^n x_1 y - \hat{\beta}_2\sum_i^n x_2 y}{n-3}$	**(10.54)**

Τα τυπικά σφάλματα των συντελεστών β_i και συγκεκριμένα για την περίπτωση όπου έχουμε δύο ανεξάρτητες μεταβλητές είναι:

$s^2_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_i^n (X_1-\bar{X}_1)^2 - \frac{\left(\sum_i^n (X_1-\bar{X}_1)(X_2-\bar{X}_2)\right)^2}{\sum_i^n (X_2-\bar{X}_2)^2}}} = \frac{s}{\sqrt{\sum_i^n x_1^2 - \frac{\left(\sum_i^n x_1 x_2\right)^2}{\sum_i^n x_2^2}}}$ και $s^2_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum_i^n (X_2-\bar{X}_2)^2 - \frac{\left(\sum_i^n (X_1-\bar{X}_1)(X_2-\bar{X}_2)\right)^2}{\sum_i^n (X_1-\bar{X}_1)^2}}} = \frac{s}{\sqrt{\sum_i^n x_2^2 - \frac{\left(\sum_i^n x_1 x_2\right)^2}{\sum_i^n x_1^2}}}$	**(10.55)**

Επομένως, τόσο για τον έλεγχο της υπόθεσης ότι ο κάθε συντελεστής β_2 έχει συγκεκριμένη τιμή, όσο και για την κατασκευή του διαστήματος εμπιστοσύνης που θα τον περιλαμβάνει με δεδομένη πιθανότητα, θα χρησιμοποιήσουμε τη στατιστική z ή την t-student, ανάλογα

με το μέγεθος του δείγματος.

Για τους ελέγχους υποθέσεων

$$H_0: \beta_i = \beta_i^* \quad \text{ή} \quad H_0: \beta_i = 0$$
$$H_r: \beta_i \neq \beta_i^* \quad\quad H_r: \beta_i \neq 0$$

και όταν το δείγμα είναι μικρό, χρησιμοποιούμε τη στατιστική:

$t = \dfrac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}}} \quad (v = n - k + 1)$	**(10.56)**

Επομένως, απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας α, αν:

$$t < -t_{n-k-1, \alpha/2} \text{ ή } t > t_{n-k-1, \alpha/2}$$

Το διαστήματος εμπιστοσύνης κάθε συντελεστή β_i είναι:

$\hat{\beta}_i - t_{\alpha/2} \cdot s_{\hat{\beta}} \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2} \cdot s_{\hat{\beta}}$	**(10.57)**

11. Εκπαιδευτική Ενότητα

- **Μη Παραμετρικοί Έλεγχοι**

ΕΚΠΑΙΔΕΥΤΙΚΟΙ ΣΤΟΧΟΙ

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατανοεί τους μη παραμετρικούς ελέγχους.
- Να εφαρμόζει τους κατάλληλους ελέγχους.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

- Chi-Square
- Runs Test
- Kolmogorov-Smirnov ενός δείγματος
- Mann-Whitney
- Wilcoxon
- Kruskal-Wallis

11.1 Εισαγωγή

Κατά την ανάλυση της στατιστικής συμπερασματολογίας, ασχοληθήκαμε με την εκτίμηση και τον έλεγχο παραμέτρων που προσδιόριζαν ένα ή περισσότερους πληθυσμούς. Ειδικότερα, για τις ανάγκες της εκτίμησης και του ελέγχου, χρησιμοποιήθηκαν στατιστικά κριτήρια τα οποία για την εφαρμογή τους, απαιτούσαν την ικανοποίηση υποθέσεων σχετικά με συγκεκριμένες παραμέτρους του πληθυσμού και τη μορφή της κατανομής τους.

Όμως, υπάρχουν περιπτώσεις όπου τα στοιχεία τα οποία διαθέτουμε για ανάλυση, δεν γνωρίζουμε από ποια κατανομή προήλθαν και επομένως, η τυχόν εφαρμογή γνωστών παραμετρικών κριτηρίων (χ^2 test, t-test κ.λ.π.) θα έδινε εσφαλμένα αποτελέσματα. Στην περίπτωση αυτή, μπορούμε να χρησιμοποιήσουμε τα λεγόμενα μη παραμετρικά κριτήρια, τα οποία δεν ελέγχουν υποθέσεις αναφορικά με συγκεκριμένες παραμέτρους του πληθυσμού και το σημαντικότερο, δεν υποθέτουν ότι οι τιμές των παραμέτρων αυτών πρέπει να ακολουθούν την κανονική κατανομή.

Όπως και τα παραμετρικά κριτήρια, έτσι και τα μη παραμετρικά, έχουν πλεονεκτήματα και μειονεκτήματα. Τα κυριότερα πλεονεκτήματα και μειονεκτήματα των μη παραμετρικών κριτηρίων είναι:

1. Πλεονεκτήματα

- Δεν προϋποθέτουν ότι τα στοιχεία ενός διαθέσιμου δείγματος προέρχονται από κανονικό πληθυσμό.
- Εφαρμόζονται ακόμα και όταν το μέγεθος του δείγματος είναι πολύ μικρό (n=6), εκτός και αν γνωρίζουμε ακριβώς την κατανομή του πληθυσμού.
- Μπορούν να εφαρμοστούν σε δεδομένα μιας ποιοτικής μεταβλητής με περισσότερες από δύο κατηγορίες.
- Μπορούν να εφαρμοστούν ακόμα και σε δεδομένα τα οποία είναι διατάξιμα (π.χ. προτιμήσεις καταναλωτών σχετικά με την ποιότητα ενός προϊόντος).
- Μπορούμε να κάνουμε έλεγχο υποθέσεων για πολλούς διαφορετικούς πληθυσμούς.
- Γενικώς, τα μη παραμετρικά κριτήρια εφαρμόζονται εύκολα.

2. Μειονεκτήματα

- Είναι λιγότερο ισχυρά σε σχέση με τα παραμετρικά.
- Δεν βασίζονται στα πραγματικά δεδομένα μιας έρευνας αλλά στην ιεραρχική τους διάταξη.
- Δεν μπορούν να χρησιμοποιηθούν για έλεγχο της αλληλεπίδρασης στην ανάλυση της διακύμανσης.
- Δεν μπορούμε να υπολογίσουμε εύκολα διαστήματα εμπιστοσύνης.

Στο κεφάλαιο αυτό, θα αναφερθούμε στα σημαντικότερα μη παραμετρικά κριτήρια τα οποία αφορούν ελέγχους καλής προσαρμογής μιας κατανομής και ελέγχους για τη διαφορά της θέσης δύο κατανομών.

11.2 Κριτήριο ελέγχου των Kolmogorov – Smirnov (K-S)

Με τον έλεγχο των Kolmogorov-Smirnov (K-S), προσπαθούμε να προσδιορίσουμε αν τα στοιχεία ενός δείγματος, προέρχεται από πληθυσμό στον οποίο ακολουθούν μια δεδομένη θεωρητική κατανομή (κανονική, διωνυμική, Poisson κ.λπ.). Για το λόγο αυτό, ονομάζεται και έλεγχος καλής προσαρμογής (test of goodness of fit).

Εν αντιθέσει με τον έλεγχο καλής προσαρμογής με το στατιστικό χ^2-test, ο έλεγχος προσαρμογής με το κριτήριο των Kolmogorov-Smirnov (K-S) έχει τα ακόλουθα πλεονεκτήματα και μειονεκτήματα:

Πλεονεκτήματα: α) Μπορεί να εφαρμοστεί και σε πολύ μικρά δείγματα. β) Δεν απαιτείται ομαδοποίηση των στοιχείων και γ) Δεν απαιτείται συγχώνευση των στοιχείων όταν οι θεωρητικές συχνότητες είναι μικρές.

Μειονεκτήματα: α) Χρησιμοποιείται όταν έχουμε παρατηρήσεις μιας συνεχούς τυχαίας μεταβλητής και β) Απαιτείται να γνωρίζουμε τον μέσο και την τυπική απόκλιση του πληθυσμού.

Σε περίπτωση που δεν γνωρίζουμε τον μέσο και την τυπική απόκλιση του πληθυσμού, τότε εφαρμόζουμε το κριτήριο ελέγχου του Lilliefors που αποτελεί επέκταση του κριτηρίου ελέγχου καλής προσαρμογής των K-S, οπότε ο μέσος και η τυπική απόκλιση εκτιμώνται από τα στοιχεία του δείγματος. Κατά τ' άλλα, για τον έλεγχο χρησιμοποιούμε την ίδια διαδικασία με αυτή των K-S.

Ειδικότερα, με τον έλεγχο των K-S συγκρίνουμε την αθροιστική συνάρτηση συχνοτήτων ενός δείγματος $F_S(X)$ με την αθροιστική συνάρτηση συχνοτήτων μιας θεωρητικής κατανομής $F_p(X)$ του πληθυσμού.

Διατύπωση υποθέσεων

$\mathrm{H}_0:\ F(X) = F_p(X)$ για όλες τις τιμές του X ή

Η κατανομή είναι κανονική, διωνυμική κ.λ.π.

$\mathrm{H}_\varepsilon:\ F(X) \neq F_p(X)$ για ένα τουλάχιστον X ή

Κάποια άλλη

Κριτήριο ελέγχου

Οι παραπάνω υποθέσεις, ελέγχονται με την ποσότητα

$$D = \max_{\forall X}\{|F_S(X) - F_p(X)|\}$$

Απόφαση

- Αν $D \geq D_{a,n}$, τότε απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική $H_ε$.
- Αν $D < D_{a,n}$, τότε αποδεχόμαστε τη μηδενική υπόθεση H_0 και απορρίπτουμε την εναλλακτική $H_ε$.

Η τιμή του $D_{a,n}$, προκύπτει από τους Στατιστικούς Πίνακες, για μέγεθος δείγματος n και επίπεδο σημαντικότητας α.

Παράδειγμα 11.1

Από έρευνα εργαζομένων στον ξενοδοχειακό κλάδο, προέκυψε ότι η ωριαία αμοιβή (X_i σε €) 34 εργαζομένων έχει ως εξής:

5,5	3,9	3,5	4,6	4,3	3,5	3,5	3,6	3,6	5,0	4,7	4,5	5,5	4,4	4,9	5,4	4,5
4,8	4,8	4,8	4,6	4,2	4,4	4,2	4,7	7,9	4,7	4,7	4,2	9,5	5,9	6,5	8,5	10,5

Να ελεγχθεί σε επίπεδο σημαντικότητας α, αν τα στοιχεία αυτά προέρχονται από κανονική κατανομή με μέσο 5,11 € και τυπική απόκλιση 1,66 €.

Απάντηση:

Στην περίπτωση αυτή, ο έλεγχος υποθέσεων διατυπώνεται ως εξής:

H_0 : Η μεταβλητή X ακολουθεί την κανονική κατανομή με μ=5,11 και σ=1,66

$H_ε$: Η μεταβλητή X ακολουθεί άλλη κατανομή

Η διαδικασία που ακολουθούμε για να ελέγξουμε τις υποθέσεις αυτές με το κριτήριο των K-S, είναι:

Βήμα 1ο: Ταξινομούμε τα στοιχεία της μεταβλητής του δείγματος κατά αύξουσα διάταξη, βρίσκουμε τις απόλυτες συχνότητες f_i και υπολογίζουμε τις δεξιόστροφες απόλυτες και σχετικές αθροιστικές συχνότητες F_i και $F_S(X_i) = F_i / n$.

Βήμα 2ο: Με μ=5,11 και σ=1,66, υπολογίζουμε τις τυπικές τιμές $Z_i = \frac{(X_i - \mu)}{\sigma}$.

Βήμα 3ο: Υπολογίζουμε τις αθροιστικές θεωρητικές συχνότητες $F_P(X_i) = P(Z \leq Z_i)$.

Βήμα 4ο: Υπολογίζουμε τις διαφορές $D_i = |F_S(X_i) - F_P(X_i)|$ και στη συνέχεια εντοπίζουμε τη μέγιστη τιμή αυτών ($D = \max D_i$). Στην περίπτωσή μας (στήλη 7)

$\max D_i = D = 0{,}275$

Πίνακας 11.1 Πίνακας ελέγχου με το κριτήριο των Kolmogorov-Smirnov

Ωριαία αμοιβή (X_i)	f_i	F_i	$F_S(X_i) = \frac{F_i}{n}$	$Z_i = \frac{(X_i-\mu)}{\sigma}$	$F_P(X_i) = P(Z \le Z_i)$	$D_i = \lvert F_S(X_i) - F_P(X_i) \rvert$
1	**2**	**3**	**4**	**5**	**6**	**7**
3,5	3	3	0,088	-0,97	0,166	-0,078
3,6	2	5	0,147	-0,91	0,181	-0,034
3,9	1	6	0,176	-0,73	0,233	-0,056
4,2	3	9	0,265	-0,55	0,291	-0,026
4,3	1	10	0,294	-0,49	0,312	-0,018
4,4	2	12	0,353	-0,43	0,334	0,019
4,5	2	14	0,412	-0,37	0,356	0,056
4,6	2	16	0,471	-0,31	0,378	0,092
4,7	4	20	0,588	-0,25	0,401	0,187
4,8	3	23	0,676	-0,19	0,425	0,252
4,9	1	24	0,706	-0,13	0,448	0,258
5,0	1	25	0,735	-0,07	0,460	**0,275**
5,4	1	26	0,765	0,17	0,571	0,193
5,5	2	28	0,824	0,23	0,595	0,229
5,9	1	29	0,853	0,48	0,684	0,169
6,5	1	30	0,882	0,84	0,800	0,083
7,9	1	31	0,912	1,68	0,955	-0,043
8,5	1	32	0,941	2,04	0,979	-0,038
9,5	1	33	0,971	2,64	0,996	-0,025
10,5	1	34	1,000	3,25	0,999	0,001

Βήμα 5ο: Για α=0,05 και n=34, υπολογίζουμε την τιμή $D_{a,n} = D_{0.05,34}$.

Βήμα 6ο: Επειδή $D = 0{,}275 > D_{0.05,34} = 0{,}227$, απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική, δηλαδή ότι τα στοιχεία του δείγματος δεν ακολουθούν την κανονική κατανομή.

11.3 Κριτήριο ελέγχου του Wilcoxon (T) των σημασμένων διαβαθμίσεων

Ο έλεγχος με το κριτήριο του Wilcoxon εφαρμόζεται όταν έχουμε δύο εξαρτημένα δείγματα τα οποία ακολουθούν οποιαδήποτε κατανομή πληθυσμού και έχουν ζευγαρωτές παρατηρήσεις. Επιπλέον, η κλίμακα μέτρησης των δεδομένων πρέπει να είναι ιεραρχική. Σύμφωνα με τον έλεγχο, η μηδενική υπόθεση εξειδικεύεται ως υπόθεση της ισότητας των μέσων ή ως ισότητα των διαμέσων, γι' αυτό ονομάζεται και έλεγχος για τη διαφορά της

θέσης (μέσο ή διάμεσο) δύο κατανομών πληθυσμού.

Διατύπωση υποθέσεων

$H_0: \mu_X = \mu_Y$

$H_1:$ i) $\mu_X \neq \mu_Y$

ii) $\mu_X > \mu_Y$

iii) $\mu_X < \mu_Y$

Κριτήριο ελέγχου

α) Μικρά δείγματα (n≤25)

Αν T η διαβάθμιση των απολύτων τιμών των διαφορών των κατά ζεύγη παρατηρήσεων των μεταβλητών X_i και Y_i $(d_i = X_i - Y_i)$ και T^+, T^- και X το άθροισμα των ψηφίων των θετικών και αρνητικών διαβαθμίσεων (ισχύει $T = T^+ + T^- = \frac{n(n+1)}{2}$), τότε το στατιστικό T του ελέγχου είναι για μεν τον δίπλευρο έλεγχο $T = \min(T^-, T^+)$, για δε τον δεξιόπλευρο και αριστερόπλευρο έλεγχο το T^- και T^+ αντίστοιχα.

Στους στατιστικούς πίνακες δίνεται η κριτική τιμή T_W για μέγεθος δείγματος n και επίπεδο σημαντικότητας α (δίπλευρος έλεγχος) και $\alpha/2$ (μονόπλευρος έλεγχος).

β) Μεγάλα δείγματα (n>25)

Όταν το μέγεθος του δείγματος είναι μεγάλο, τότε ακολουθείται ο έλεγχος της κατανομής

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$	**(11.1)**

Σε περίπτωση που έχουμε ισοψηφίες διαβάθμισης, τότε η διακύμανση της κατανομής Τ υπολογίζεται από τη σχέση:

$$\sigma_T^2 = \frac{1}{24}\left(n(n+1)(2n+1) - \frac{1}{2}\sum_J t_j\left(t_j^2 - 1\right)\right) \quad \textbf{(11.2)}$$

Όπου: j = ομάδες ισοψηφιών και t_j = αντίστοιχες διαφορές σε κάθε ομάδα.

Απόφαση

α) Μικρά δείγματα (n≤25)

Απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική $H_ε$ σε επίπεδο σημαντικότητας α, αν για τα παραπάνω τρία είδη ελέγχου ισχύει αντίστοιχα:

i) T^+ ή $T^- < T_W$, ii) $T^- < T_W$ και iii) $T^+ < T_W$.

β) Μεγάλα δείγματα (n>25)

Απορρίπτουμε τη μηδενική υπόθεση H_0 σε επίπεδο σημαντικότητας α αν ισχύει για τα παραπάνω τρία είδη ελέγχου αντίστοιχα:

i) $z_W = \dfrac{T - \mu_T}{\sigma_T} < -z_{1-a/2}$, ii) $z_W = \dfrac{T^- - \mu_T}{\sigma_T} < -z_{1-a}$ και iii) $z_W = \dfrac{T^+ - \mu_T}{\sigma_T} < -z_{1-a}$.

Παράδειγμα 11.2

Από έρευνα 20 εργαζομένων (10 άνδρες και 10 γυναίκες) σε μία επιχείρηση που έχουν το ίδιο επίπεδο εκπαίδευσης, προέκυψε ότι η ωριαία αμοιβή τους (σε €) έχει ως εξής:

Ωριαία αμοιβή ανδρών	6	7	2	8	4	5	3	10	15	7
Ωριαία αμοιβή γυναικών	5	5	5	4	9	4	7	4	8	9

Να ελεγχθεί σε επίπεδο σημαντικότητας α, αν η μέση ωριαία αμοιβή των ανδρών διαφέρει από την αντίστοιχη των γυναικών.

Απάντηση

Στην περίπτωση αυτή, ο έλεγχος υποθέσεων διατυπώνεται ως εξής:

$H_0: \mu_{ανδρών} = \mu_{γυναικών}$ (Δεν διαφέρουν οι μέσες αμοιβές μεταξύ των δύο φύλων)

$H_ε: \mu_{ανδρών} \neq \mu_{γυναικών}$ (Διαφέρουν οι μέσες αμοιβές μεταξύ των δύο φύλων)

Η διαδικασία ελέγχου των παραπάνω υποθέσεων με το κριτήριο του Wilcoxon (T), ακολουθεί τα εξής βήματα:

Βήμα 1ο: Βρίσκουμε τις διαφορές d_i ανά ζεύγος (αφαιρούμε πάντα από την ίδια κατεύθυνση) των τιμών των μεταβλητών. Δηλαδή:

Ωριαία αμοιβή ανδρών (X_i)	6	7	2	8	4	5	3	10	15	7
Ωριαία αμοιβή γυναικών (Y_i)	5	5	5	4	9	4	7	4	8	9
Διαφορές ωρομισθίων (d_i)	1	2	-3	4	-5	1	-4	6	7	-2

Βήμα 2ο: Ταξινομούμε τις παραπάνω διαφορές (d_i) κατά αύξουσα διάταξη απολύτων τιμών, αγνοώντας τις μηδενικές διαφορές. Και στη συνέχεια διαβαθμίζουμε τις μη μηδενικές απόλυτες τιμές των διαφορών κατά σειρά μεγέθους (1 στη μικρότερη απόλυτη διαφορά κ.ο.κ.). Σε περίπτωση που έχουμε ισοψηφίες διαφορών (δηλαδή απόλυτες διαφορές d_i με την ίδια τιμή), τότε ως αριθμός διαβάθμισης τίθεται ο μέσος όρος των θέσεων που κατέχουν από την αύξουσα ταξινόμησή τους. Δηλαδή:

Αύξουσα ταξινόμηση (d_i)	1	1	2	2	3	4	4	5	6	7
Διαβάθμιση διαφορών (d_i)	1,5	1,5	3,5	3,5	5	6,5	6,5	8	9	10

Βήμα 3ο: Σε κάθε διαβάθμιση, σημειώνουμε το πρόσημο της αντίστοιχης διαφοράς που προέκυψε από το 1ο βήμα.

Πρόσημο διαβάθμισης διαφορών (d_i)	+1,5	+1,5	-3,5	+3,5	-5	6,5	-6,5	+8	+9	-10

Βήμα 4ο: Υπολογίζουμε το άθροισμα των θετικών και αρνητικών διαβαθμίσεων, T^+ και αντίστοιχα.

Θετικές διαβαθμίσεις T^+ = 1,5+1,5+3,5+6,5+8+9 = 30

Αρνητικές διαβαθμίσεις T^- = 3,5+5+6,5+10 = 25

Επομένως ισχύει: $T = T^+ + T^- = \dfrac{n(n+1)}{2} = 30+25 = \dfrac{10(10+1)}{2} = 55$

Βήμα 5ο: Αν το δείγμα είναι μικρό (n=10≤25), τότε για τον έλεγχο χρησιμοποιούμε το στατιστικό T που η τιμή του προκύπτει από το μικρότερο άθροισμα των θετικών και αρνητικών διαβαθμίσεων, δηλαδή:

$$T = \min(T^-, T^+) = \min(25{,}30) = 25 .$$

Βήμα 6ο: Σε επίπεδο σημαντικότητας α=0,05 και n=10, προκύπτει από τους στατιστικούς πίνακες ότι $T_W = 8$. Επομένως, αφού, $T = T^- = 25 > T_W = 8$ αποδεχόμαστε τη μηδενική υπόθεση που σημαίνει ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μέσων ωριαίων αμοιβών ανδρών και γυναικών.

11.4 Κριτήριο ελέγχου των Mann-Whitney (U)

Με το κριτήριο αυτό, ελέγχουμε τη διαφορά της θέσης δύο κατανομών πληθυσμού και χρησιμοποιείται όταν δεν ισχύει η υπόθεση ότι οι κατανομές πληθυσμού είναι κανονικές. Για να εφαρμόσουμε το κριτήριο αυτό, θα πρέπει τα δείγματα να είναι τυχαία και ανεξάρτητα, οι τυχαίες μεταβλητές να είναι συνεχείς, οι κατανομές των δύο δειγμάτων να έχουν την ίδια μορφή και η κλίμακα μέτρησης των δεδομένων ιεραρχική. Το κριτήριο των MannWhitney (U) είναι το αντίστοιχο του t-test για ανεξάρτητα δείγματα. Για δύο τυχαία ανεξάρτητα δείγματα X και Y μεγέθους n_x και n_y αντίστοιχα, η διατύπωση των υποθέσεων εξειδικεύεται ως ακολούθως:

Διατύπωση υποθέσεων

$$H_0 : \mu_X = \mu_Y$$

$$H_1 : \text{i)}\ \mu_X \neq \mu_Y$$

$$\text{ii)}\ \mu_X > \mu_Y$$

$$\text{iii)}\ \mu_X < \mu_Y$$

Κριτήριο ελέγχου

α) Πολύ μικρά δείγματα ($n_x \leq n_y \leq 8$) και μικρά δείγματα $8 \leq \max\{n_X, n_Y\} \leq 20$

Αν T η από κοινού διαβάθμιση των παρατηρήσεων των μεταβλητών X και Y και T_x, T_y το άθροισμα των ψηφίων των διαβαθμίσεων του πρώτου και δεύτερου δείγματος αντίστοιχα, τότε:

$U_X = n_X n_Y + \dfrac{n_X(n_X+1)}{2} - T_X \text{ και } U_Y = n_X n_Y + \dfrac{n_Y(n_Y+1)}{2} - T_Y$	**(11.3)**

Το στατιστικό U του ελέγχου είναι για μεν τον δίπλευρο έλεγχο U=min(U_X, U_Y) για δε τον δεξιόπλευρο και αριστερόπλευρο έλεγχο το U_X και U_Y αντίστοιχα.

β) Μεγάλα δείγματα ($\max\{n_X, n_Y\} > 20$)

Σ' αυτή την περίπτωση, όταν ισχύει η μηδενική υπόθεση H_0 η τυχαία μεταβλητή U προσεγγίζει την κανονική κατανομή. Δηλαδή, ισχύει:

$U \sim N\left(\frac{n_X n_Y}{2}, \frac{n_X n_Y (n_X + n_Y + 1)}{12}\right)$ και $Z_U = \frac{U - \mu_U}{\sigma_U}$	**(11.4)**

Σε περίπτωση που έχουμε ισοψηφίες διαβάθμισης, τότε η διακύμανση της κατανομής U υπολογίζεται από τη σχέση

$\sigma_U^2 = \frac{n_X n_Y}{12}\left(n_X + n_Y + 1 - \frac{\sum_j t_j (t_j^2 - 1)}{(n_X + n_Y)(n_X + n_Y + 1)}\right)$	**(11.5)**

Όπου: j = ομάδες ισοψηφιών και t_j = αντίστοιχες διαφορές σε κάθε ομάδα.

Απόφαση

α) Πολύ μικρά δείγματα ($n_x \leq n_y \leq 8$)

Απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική H_ε σε επίπεδο σημαντικότητας α, αν για τα παραπάνω τρία είδη ελέγχου ισχύει αντίστοιχα:

i) $\min(U_X, U_Y) < \frac{a}{2}$, ii) $U_Y < a$ και iii) $U_X < a$.

β) Μικρά δείγματα ($8 \leq \max\{n_x, n_y\} \leq 20$)

Απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική H_ε σε επίπεδο σημαντικότητας α, αν για τα παραπάνω τρία είδη ελέγχου ισχύει αντίστοιχα:

i) $\min(U_X, U_Y) < \frac{a}{2}$, ii) $U_Y < a$ και iii) $U_X < a$.

γ) Μεγάλα δείγματα ($\max\{n_x, n_y\}>20$)

Απορρίπτουμε τη μηδενική υπόθεση H_0 σε επίπεδο σημαντικότητας H_ε αν ισχύει για τα παραπάνω τρία είδη ελέγχου αντίστοιχα:

i) $|z_U| = \left| \dfrac{U_X \text{ ή } U_Y - \mu_U}{\sigma_U} \right| > z_{1-\alpha/2}$, ii) $z_U = \dfrac{U_X \text{ ή } U_Y - \mu_U}{\sigma_U} > z_a$ και

iii) $z_U = \dfrac{U_X \text{ ή } U_Y - \mu_U}{\sigma_U} < -z_a$.

Παράδειγμα 11.3

Για τις ανάγκες μιας έρευνας αγοράς, συλλέχθηκαν οι ακόλουθες μηνιαίες διαφημιστικές δαπάνες (σε €) ενός έτους που πραγματοποίησαν δύο επιχειρήσεις που δραστηριοποιούνται στον κλάδο της γαλακτοβιομηχανίας:

Επιχείρηση Α	115	125	110	130	122	118	125	129	132	135	124	137
Επιχείρηση Β	118	127	102	135	110	125	125	130	135	132	125	135

Να ελεγχθεί σε επίπεδο σημαντικότητας $a=0.10$, αν η μέση μηνιαία διαφημιστική δαπάνη που πραγματοποίησε η επιχείρηση Α διαφέρει από αυτή της Β.

Απάντηση

Στην περίπτωση αυτή, ο έλεγχος υποθέσεων διατυπώνεται ως εξής:

H_0 : $\mu_A = \mu_B$ (Δεν διαφέρουν οι μέσες διαφημιστικές δαπάνες των επιχειρήσεων Α και Β)

H_ε : $\mu_A \neq \mu_B$ (Διαφέρουν οι μέσες διαφημιστικές δαπάνες των επιχειρήσεων Α και Β)

Η διαδικασία ελέγχου των υποθέσεων με το κριτήριο των Mann-Whitney (U), ακολουθεί τα εξής βήματα:

Βήμα 1^{ο}: Προσδιορίζουμε τα μεγέθη των δύο δειγμάτων (n_x και n_y), τα ενώνουμε σε ένα και στη συνέχεια ταξινομούμε τα στοιχεία του ενιαίου δείγματος κατά αύξουσα διάταξη λαμβάνοντας υπόψη και το πρόσημό τους.

Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Α+Β	102	110	110	115	118	118	122	124	125	125	125	125	125	127	129	130	130	132	132	135	135	135	135	137

Βήμα 2ο: Διαβαθμίζουμε (T) τα στοιχεία του ενιαίου δείγματος αντιστοιχώντας τον αριθμό 1 στο μικρότερο στοιχείο έως την τιμή n (όπου $n=n_x+n_y$). Σε περίπτωση που έχουμε ισοψηφίες στοιχείων, τότε ως αριθμός διαβάθμισης τίθεται ο μέσος όρος των θέσεων που κατέχουν αυτά από την αύξουσα ταξινόμησή τους.

Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Μέθοδοι Α & Β	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Διαβάθμιση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Βήμα 3ο: Από την ενιαία διαβάθμιση, υπολογίζουμε το άθροισμα των διαβαθμίσεων T_x και T_y, που αφορούν τα στοιχεία των δειγμάτων n_x, n_y αντίστοιχα.

Μέθοδος Α	1	2	3	5	9	10	12	13	14	15	T_X
Διαβ/ση	1	2	3	5	9	10	12	13	14	15	84

Μέθοδος Β	4	6	7	8	11	16	17	18	19	20	T_Y
Διαβ/ση	4	6	7	8	11	16	17	18	19	20	126

Ισχύει, $T = T_X + T_Y = 145 + 155 = \dfrac{(n_X + n_Y)(n_X + n_Y + 1)}{2} = \dfrac{(12+12)(12+12+1)}{2} = 300$

Βήμα 4ο: Για κάθε δείγμα, υπολογίζουμε τις στατιστικές U_X και U_Y:

$$U_X = n_X n_Y + \frac{n_X(n_X+1)}{2} - T_X = 12\cdot 12 + \frac{12(12+1)}{2} - 145 = 77 \text{ και}$$

$$U_Y = n_X n_Y + \frac{n_Y(n_Y+1)}{2} - T_Y = 12\cdot 12 + \frac{12(12+1)}{2} - 155 = 67$$

Επίσης, προκύπτει ότι ισχύει $U_X + U_Y = n_X \cdot n_Y = 144$.

Βήμα 5ο: Επειδή τα δείγματα είναι μικρά ($8 \le \max\{n_X, n_Y\} = 12 \le 20$), για τον έλεγχο χρησιμοποιούμε το στατιστικό U. Επομένως, $U = \min(U_X, U_Y) = \min(77, 67) = 67$.

Βήμα 6^{ο}: Επειδή ο έλεγχος είναι δίπλευρος και 0,10/2=0,05, n_x=10 και n_y=10, προκύπτει από τους στατιστικούς πίνακες ότι U_{CR}=27. Επομένως, αφού U=67> U_{CR} =27, αποδεχόμαστε τη μηδενική υπόθεση που σημαίνει ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μέσων διαφημιστικών δαπανών των επιχειρήσεων Α και Β.

11.5 Κριτήριο έλεγχου του Wilcoxon (W)

Για να εφαρμοστεί ο έλεγχος αθροίσματος διατάξεων του Wilcoxon (W) (όπως και σ' αυτόν των Mann-Whitney), θα πρέπει τα δείγματα να είναι τυχαία και ανεξάρτητα, οι τυχαίες μεταβλητές να είναι συνεχείς, οι κατανομές των δύο δειγμάτων να έχουν την ίδια μορφή και η κλίμακα μέτρησης των δεδομένων ιεραρχική (ordinal scale). Το κριτήριο του Wilcoxon (W) είναι το αντίστοιχο του t-test για ανεξάρτητα δείγματα και είναι το ίδιο ισχυρό αλλά περισσότερο διαδεδομένο από τον έλεγχο των Mann-Whitney. Αν X και Y δύο τυχαία ανεξάρτητα δείγματα μεγέθους n_X και n_y αντίστοιχα, τότε, η διατύπωση των υποθέσεων, εξειδικεύεται ως ακολούθως:

Διατύπωση υποθέσεων

$$H_0: \mu_X = \mu_Y$$

$$H_\tau: \text{i)}\ \mu_X \neq \mu_Y$$

$$\text{ii)}\ \mu_X > \mu_Y$$

$$\text{iii)}\ \mu_X < \mu_Y$$

Κριτήριο ελέγχου

α) μικρά δείγματα (n_X, $n_y \leq 10$)

Αν T η από κοινού διαβάθμιση των παρατηρήσεων των μεταβλητών X και Y και T_X, T_y το άθροισμα των ψηφίων των διαβαθμίσεων του πρώτου και δεύτερου δείγματος αντίστοιχα (ισχύει $T = T_X + T_Y = \frac{(n_X+n_Y)(n_X+n_Y+1)}{2}$), τότε μπορούμε να ορίσουμε την τυχαία μεταβλητή $T = n_X(n_X+n_X+1) - T_X$ ή $T = \min(T_X, T_Y)$ ανάλογα αν ισχύει $n_X<n_y$ και $n_X= n_y$ αντίστοιχα. Το στατιστικό W του ελέγχου είναι για μεν τον δίπλευρο έλεγχο $W = \min(T_X, T_Y)$ ή $W = \min\{T, \min(T_X, T_Y)\}$ ανάλογα αν ισχύει $n_X< n_y$ και $n_X=n_y$αντίστοιχα, για δε τον δεξιόπλευρο και αριστερόπλευρο έλεγχο το $W = T_X$ και $W = T_X$.

β) Μεγάλα δείγματα (n_x, n_y>10)

Σ' αυτή την περίπτωση, όταν ισχύει η μηδενική υπόθεση H_0 το W προσεγγίζει την κανονική κατανομή. Δηλαδή, ισχύει:

$W \sim N\left(\frac{n_X(n_X+n_Y+1)}{2}, \frac{n_X n_Y(n_X+n_Y+1)}{12}\right)$ και $Z_W = \frac{W-\mu_W}{\sigma_W}$	**(11.6)**

Σε περίπτωση που έχουμε ισοψηφίες διαβάθμισης, τότε η διακύμανση της κατανομής W υπολογίζεται από τη σχέση:

$\sigma_W^2 = \frac{n_X n_Y}{12}\left(n_X+n_Y+1-\frac{\sum_j t_j(t_j^2-1)}{(n_X+n_Y)(n_X+n_Y+1)}\right)$	**(11.7)**

Όπου: j = ομάδες ισοψηφιών και t_j = αντίστοιχες διαφορές σε κάθε ομάδα.

Απόφαση

α) Πολύ μικρά δείγματα (n_x, n_y≤10)

Απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική H_ε σε επίπεδο σημαντικότητας α, αν για τα παραπάνω τρία είδη ελέγχου ισχύει αντίστοιχα:

i) $W \geq w_{\alpha/2}$ ή $W \leq -w_{\alpha/2}$, ii) $W \geq w_{\alpha/2}$ και iii) $W \leq -w_{\alpha/2}$.

β) Μεγάλα δείγματα (n_x, n_y>10)

Απορρίπτουμε τη μηδενική υπόθεση H_0 σε επίπεδο σημαντικότητας α αν ισχύει για τα παραπάνω τρία είδη ελέγχου αντίστοιχα:

i) $|Z_W| = \left|\frac{W-\mu_W}{\sigma_W}\right| > z_{\alpha/2}$, ii) $Z_W \geq z_\alpha$ και iii) $Z_W < -z_\alpha$.

Παράδειγμα 11.2

Ο γενικός διευθυντής μιας μεγάλης ναυτιλιακής εταιρείας, επέλεξε 20 υπαλλήλους του απόφοιτους τριτοβάθμιας για να παρακολουθήσουν ένα σεμινάριο κατάρτισης διάρκειας 120 ωρών σε θέματα management και marketing. Για το σκοπό αυτό, χώρισε τυχαία τους 20 εργαζομένους σε δύο τμήματα των 10 ατόμων το καθένα και αποφάσισε να χρησιμοποιηθούν δύο μέθοδοι κατάρτισης. Στο πρώτο τμήμα χρησιμοποιήθηκε η «παραδοσιακή μέθοδος», ενώ στο δεύτερο μια νέα ''πειραματική μέθοδος''. Μετά το πέρας της κατάρτισης και αφού παρήλθε διάστημα 6 μηνών εργασίας, ο γενικός διευθυντής βαθμολόγησε την απόδοσή τους με κλίμακα από το 1 (ο χειρότερος) έως το 20 (ο καλύτερος) και προέκυψαν τα ακόλουθα αποτελέσματα:

Παραδοσιακή μέθοδος κατάρτισης	1	2	3	5	9	10	12	13	14	15
Πειραματική μέθοδος κατάρτισης	4	6	7	8	11	16	17	18	19	20

Να ελεγχθεί σε επίπεδο σημαντικότητας α=0,05, αν διαφέρει η μέση απόδοση των υπαλλήλων μεταξύ των δύο μεθόδων κατάρτισης που παρακολούθησαν.

Απάντηση:

Στην περίπτωση αυτή, ο έλεγχος υποθέσεων διατυπώνεται ως εξής:

$$H_0: \mu_{ΠΑΡ.} = \mu_{ΠΕΙΡ.} \text{ (Δεν επηρεάστηκε η απόδοση των υπαλλήλων)}$$

$$H_r: \mu_{ΠΑΡ.} \neq \mu_{ΠΕΙΡ.} \text{ (Επηρεάστηκε η απόδοση των υπαλλήλων)}$$

Η διαδικασία ελέγχου των παραπάνω υποθέσεων με το κριτήριο του Wilcoxon (W), ακολουθεί τα εξής βήματα:

Βήμα 1ο: Προσδιορίζουμε τα μεγέθη των δύο δειγμάτων (n_x και n_y), τα ενώνουμε σε ένα και στη συνέχεια ταξινομούμε τα στοιχεία του ενιαίου δείγματος κατά αύξουσα διάταξη. Δηλαδή:

Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Μέθοδοι Α & Β	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Βήμα 2ο: Διαβαθμίζουμε (T) τα στοιχεία του ενιαίου δείγματος αντιστοιχώντας τον αριθμό ένα στο μικρότερο στοιχείο έως την τιμή n (όπου $n=n_x+n_y$). Σε περίπτωση που έχουμε ισοψηφίες στοιχείων, τότε ως αριθμός διαβάθμισης τίθεται ο μέσος όρος των θέσεων που κατέχουν αυτά από την αύξουσα ταξινόμησή τους.

Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Μέθοδοι Α & Β	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Διαβάθμιση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Βήμα 3°: Από την ενιαία διαβάθμιση, υπολογίζουμε το άθροισμα των διαβαθμίσεων T_x και T_y, που αφορούν τα στοιχεία των δειγμάτων n_x, n_y αντίστοιχα.

Μέθοδος A	1	2	3	5	9	10	12	13	14	15	T_x
Διαβ/ση	1	2	3	5	9	10	12	13	14	15	84

Μέθοδος B	4	6	7	8	11	16	17	18	19	20	T_y
Διαβ/ση	4	6	7	8	11	16	17	18	19	20	126

Ισχύει, $T = T_x + T_y = \frac{(n_x + n_y)(n_x + n_y + 1)}{2} = 84 + 126 = \frac{(10+10)(10+10+1)}{2} = 210$

Βήμα 4°: Αφού $n_x = n_y = 10$, τότε $T = \min(T_x, T_y) = \min(84, 126) = 84$.

Βήμα 5°: Αφού τα δείγματα είναι μικρά (n_x, n_y=10) και επιπλέον ο έλεγχος είναι δίπλευρος, χρησιμοποιούμε το στατιστικό W και η τιμή του είναι αυτή που αντιστοιχεί στη μικρότερη τιμή του στατιστικού T και της μικρότερης τιμής μεταξύ των T_x και T_y .

Δηλαδή, $W = \min\{T, \min(T_x, T_y)\} = \min\{84, \min(84, 126)\} = 84$.

Βήμα 6°: Επειδή ο έλεγχος είναι δίπλευρος και α=0,05, n_x=10 και n_y=10, προκύπτει από τους στατιστικούς πίνακες ότι $w_{\alpha/2,(10,10)}$=78 για την αριστερή ουρά της κατανομής και $w^*_{\alpha/2,(10,10)}$ =132 για τη δεξιά.

Επομένως, αφού $w_{\alpha/2,(10,10)} = 61 < W = 84 < w^*_{\alpha/2,(10,10)}$ =132, αποδεχόμαστε τη μηδενική υπόθεση που σημαίνει ότι δεν επηρεάστηκε η απόδοση των υπαλλήλων από τις διαφορετικές μεθόδους κατάρτισης που εφάρμοσε η επιχείρηση.

11.6 Κριτήριο των Kruskall–Wallis (H)

Αν η υπόθεση της κανονικότητας του F-test στην ανάλυση διακύμανσης (ANOVA) δεν ισχύει, δηλαδή ότι τα k δείγματα προέρχονται από κανονικούς πληθυσμούς, τότε μπορούμε να χρησιμοποιήσουμε τον έλεγχο των Kruskall–Wallis. Επομένως, μπορούμε να ισχυριστούμε ότι αποτελεί μια μη παραμετρική εκδοχή της ανάλυσης διακύμανσης και κατά συνέπεια έχει την ίδια σχετική ισχύ με το στατιστικό F-test της ανάλυσης διακύμανσης. Το κριτήριο ελέγχου των Kruskall–Wallis ελέγχει την ισότητα των διάμεσων τιμών των k δειγμάτων και όχι των μέσων τιμών αυτών οπότε και αποτελεί επέκταση του κριτηρίου του Wilcoxon. Με άλλα λόγια, Το κριτήριο αυτό ελέγχει αν k (k>2) ανεξάρτητα δείγματα προέρχονται από πληθυσμούς που ακολουθούν την ίδια κατανομή. Αν k=2, τότε ο έλεγχος αυτός ισοδυναμεί με τον έλεγχο των Mann-Whitney και του Wilcoxon. Για να εφαρμόσουμε τον έλεγχο των Kruskall–Wallis, θα πρέπει οι k πληθυσμοί να έχουν την ίδια μορφή, την ίδια διακύμανση και επιπλέον οι εξεταζόμενες μεταβλητές να είναι ανεξάρτητες, συνεχείς και να έχουμε τουλάχιστον πέντε παρατηρήσεις σε κάθε δείγμα $(n_j \geq 5,\ j=1,\ldots,k)$.

Διατύπωση υποθέσεων

$$H_0: M_1 = M_2 = \ldots = M_k \text{ (ή δείγματα από ίδια κατανομή)}$$

$$H_\varepsilon: M_1 \neq M_2 \neq \ldots \neq M_k \text{ (ή δείγματα από διαφορετικές κατανομές)}$$

Κριτήριο ελέγχου

Όπως και στην περίπτωση του ελέγχου με το κριτήριο των Mann-Whitney, προσδιορίζουμε κατ' αρχάς το μέγεθος των k δειγμάτων $(n_1, n_2, \ldots, n_k)$, τα συγχωνεύουμε σε ένα και κατατάσσουμε τις n $(n = n_1 + n_2 + \ldots + n_k)$ του ενιαίου δείγματος κατ' αύξουσα τάξη μεγέθους. Στη συνέχεια, διαβαθμίζουμε τα στοιχεία του ενιαίου δείγματος, ενώ σε περίπτωση που έχουμε ισοψηφίες στοιχείων, τότε ως αριθμός διαβάθμισης τίθεται ο μέσος όρος των θέσεων που κατέχουν αυτά από την αύξουσα ταξινόμησή τους.

Αν $T_1, T_2, \ldots, T_k$ το άθροισμα των τάξεων για καθένα από τα k δείγματα, τότε η στατιστική του κριτηρίου ελέγχου των Mann-Whitney ορίζεται ως εξής:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{T_j^2}{n_j} \right] - 3(n+1)$$	**(11.8)**

Όπου:
n = ο συνολικός παρατηρήσεων του ενιαίου/συνδυασμένου δείγματος
n_j = ο συνολικός παρατηρήσεων κάθε δείγματος $(j = 1, 2, \ldots, k)$
T_j = το άθροισμα των τάξεων κάθε δείγματος

T_j^2 = το τετράγωνο του αθροίσματος των τάξεων κάθε δείγματος

k = ο αριθμός των δειγμάτων

Σε περίπτωση που έχουμε ισοψηφίες διαβάθμισης, τότε το στατιστικό Η θα πρέπει να διορθωθεί κατά τον παράγοντα $C = 1 - \frac{\sum_J t_j(t_j^2 - 1)}{n(n^2+1)}$, οπότε αντί το στατιστικό Η, χρησιμοποιούμε το διορθωμένο στατιστικό:

$$H' = \frac{H}{C} = \frac{\left[\frac{12}{n(n+1)}\sum_{j=1}^{k}\frac{T_j^2}{n_j}\right] - 3(n+1)}{1 - \frac{\sum_J t_j(t_j^2-1)}{n(n^2+1)}}$$	**(11.9)**

Όπου
j = ομάδες ισοψηφιών και
t_j = αντίστοιχες διαφορές σε κάθε ομάδα.

Όταν το μέγεθος κάθε δείγματος είναι μεγαλύτερο από 5 $(n_j > 5, \forall i)$, τότε το στατιστικό Η ακολουθεί προσεγγιστικά την κατανομή χ^2 με $k-1$ βαθμούς ελευθερίας.

Απόφαση

Απορρίπτουμε τη μηδενική υπόθεση H_0 και αποδεχόμαστε την εναλλακτική H_ε σε επίπεδο σημαντικότητας α, αν Η ή $H' > \chi^2_{k-1,a}$. Δηλαδή, απορρίπτουμε την υπόθεση ότι τα δείγματα προήλθαν από τον ίδιο πληθυσμό.

Παράδειγμα 11.2

Ο διευθυντής πωλήσεων μιας αλυσίδας supermarkets θέλει να εξακριβώσει εάν η τοποθέτηση ζωοτροφών για κατοικίδια ζώα σε ράφια που βρίσκονται στο μπροστινό μέρος του supermarket, στη μέση και στο πίσω μέρος αυτού, έχει κάποια επίδραση στις πωλήσεις τους. Για το σκοπό αυτό, επιλέγει τυχαία ένα δείγμα 15 καταστημάτων της αλυσίδας, τα οποία ταξινομεί επίσης τυχαία ανά 5 με βάση τον σχεδιασμό (την τοποθέτηση του προϊόντος σε ράφια που βρίσκονται στο μπροστινό τμήμα του καταστήματος, στη μέση και στο πίσω μέρος αυτού). Τα καταστήματα, ακολουθώντας το πειραματικό σχέδιο του διευθυντή πωλήσεων, άλλαξαν (σε όποιο κατάστημα απαιτούνταν) το σημείο τοποθέτησης των προϊόντων σύμφωνα με το σχέδιο. Ο αριθμός των νοικοκυριών της περιοχής δραστηριότητας του καταστήματος και η τιμή του προϊόντος, θεωρούνται παράγοντες που δεν επηρεάζουν τις πωλήσεις του προϊόντος. Μετά την πάροδο ενός μηνός, ο διευθυντής κατέγραψε για κάθε κατάστημα τις ακόλουθες πωλήσεις (σε €):

Πωλήσεις ζωοτροφών με το προϊόν τοποθετημένο σε ράφια που βρίσκονται στο **μπροστινό μέρος** του καταστήματος **(Α)**	Πωλήσεις ζωοτροφών με το προϊόν τοποθετημένο σε ράφια που βρίσκονται στη **μεσαίο μέρος** του καταστήματος **(Β)**	Πωλήσεις ζωοτροφών με το προϊόν τοποθετημένο σε ράφια που βρίσκονται στο **πίσω μέρος** του καταστήματος **(Γ)**
8,6	3,2	4,6
7,2	2,4	6,0
5,4	2,0	4,0
6,2	1,4	2,8
5,0	1,8	2,2

Να ελεγχθεί σε επίπεδο σημαντικότητας α=0,01 αν οι διάμεσες πωλήσεις ζωοτροφών που προέρχονται μετά την τοποθέτηση των προϊόντων σε ράφια διαφορετικής θέσης στο χώρο των supermarkets διαφέρουν μεταξύ τους.

Απάντηση:

Ο έλεγχος υποθέσεων διατυπώνεται ως εξής:

$\mathrm{H}_0 : M_{(\mathrm{A})} = M_{(\mathrm{B})} = M_{(\Gamma)}$ (Δεν διαφέρουν οι διάμεσες πωλήσεις σύμφωνα με τον σχεδιασμό)

$\mathrm{H}_s : M_{(\mathrm{A})} \neq M_{(\mathrm{B})} \neq M_{(\Gamma)}$ (Διαφέρουν οι διάμεσες πωλήσεις σύμφωνα με τον σχεδιασμό)

Η διαδικασία ελέγχου των παραπάνω υποθέσεων με το κριτήριο των Kruskall–Wallis, ακολουθεί τα εξής βήματα:

Βήμα 1°: Ενώνουμε τα τρία δείγματα σε ένα και στη συνέχεια ταξινομούμε τα στοιχεία του ενιαίου δείγματος κατά αύξουσα διάταξη. Δηλαδή:

Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
(Α), (Β) & (Γ)	1,4	1,8	2,0	2,2	2,4	2,8	3,2	4,0	4,6	5,0	5,4	6,0	6,2	7,2	8,6

Βήμα 2°: Διαβαθμίζουμε (Τ) τα στοιχεία του ενιαίου δείγματος αντιστοιχώντας τον αριθμό ένα 1 στο μικρότερο στοιχείο έως την τιμή 15.

Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
(Α), (Β) & (Γ)	1,4	1,8	2,0	2,2	2,4	2,8	3,2	4,0	4,6	5,0	5,4	6,0	6,2	7,2	8,6
Διαβάθμιση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Βήμα 3°: Από την ενιαία διαβάθμιση, υπολογίζουμε το άθροισμα των διαβαθμίσεων T_x, T_y και T_z, που αφορούν τα στοιχεία των τριών δειγμάτων αντίστοιχα.

(Α)	8,6	7,2	5,4	6,2	5	T_X
Διαβάθμιση	15	14	11	13	10	63

(Β)	3,2	2,4	2	1,4	1,8	T_Y
Διαβάθμιση	7	5	3	1	2	18

(Γ)	4,6	6	4	2,8	2,2	T_Z
Διαβάθμιση	9	12	8	6	4	39

Βήμα 4°: Υπολογίζουμε τη στατιστική:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{T_j^2}{n_j} \right] - 3(n+1) = \frac{12}{15(15+1)} \left(\frac{63^2}{5} + \frac{18^2}{5} + \frac{39^2}{5} \right) - 3(15+1) = 10{,}14$$

Βήμα 5^{ο}: Επειδή ο έλεγχος είναι δίπλευρος και α=0,01, n_x=5, n_y=5 και n_z=5, προκύπτει από τους στατιστικούς πίνακες ότι $H_{0,01}=7,98$.

Επομένως, αφού $H=10,14>H_{0,01}=7,98$, απορρίπτουμε τη μηδενική υπόθεση που σημαίνει ότι οι διάμεσες πωλήσεις των καταστημάτων διαφέρουν μεταξύ τους ως προς τις τρεις διαφορετικές θέσεις τοποθέτησης του προϊόντος (δηλαδή σε ράφια διαδρόμων στο μπροστινό μέρος των καταστημάτων, στη μέση και στο πίσω μέρος αυτών).

ΕΡΩΤΗΣΕΙΣ

1) Ερώτηση:
Δώστε ένα ορισμό για τον Πληθυσμό και το Δείγμα σε μια στατιστική έρευνα.

2) Ερώτηση:
Περιγράψτε συνοπτικά τον όρο στατιστική σκέψη.

3) Ερώτηση:
Τι έχετε να παρατηρήσετε για τα παρακάτω επιλεγόμενα δείγματα:

α....Κάποιος θέλει να σχηματίσει μια ιδέα για το αποτέλεσμα των επερχόμενων βουλευτικών εκλογών. Τηλεφωνεί λοιπόν σε συγγενείς και φίλους του και τους ρωτάει σχετικά.

β....Για να εκτιμήσουμε το κατά κεφαλή εισόδημα των Ελλήνων παίρνουμε ένα δείγμα από το Κολωνάκι των Αθηνών.

4) Ερώτηση:
Για να βρούμε ποιες εκπομπές στην τηλεόραση έχουν μεγαλύτερη ακροαματικότητα αποφασίζουμε να πάρουμε δείγμα 500 τηλεθεατών. Ποιος είναι, κατά τη γνώμη σας, ο καλύτερος από τους παρακάτω τρόπους, για να πάρουμε δείγμα;

α) μόνο γυναίκες
β) μόνο άντρες
γ) άτομα από τις μεγάλες πόλεις
δ) άτομα μόνο από την επαρχία
ε) άτομα από διάφορες περιοχές

5) Ερώτηση:
Έστω $x_1, x_2, \dots, x_{11}$ ένα δείγμα με παρατηρήσεις:
7, 5, α, 2, 5, β, 8, 6, γ, 5, 3
όπου α, β, γ φυσικοί αριθμοί. Δίνεται ότι η μέση τιμή = 6.

Να βρεθούν οι τιμές των α, β, γ έτσι ώστε να ισχύει α+β+γ=25.

6) Ερώτηση:
Χαρακτηρίστε τις επόμενες προτάσεις ως σωστές ή λανθασμένες:

1. Όταν έχουμε ακραίες παρατηρήσεις, είναι προτιμότερο να χρησιμοποιούμε τη μέση τιμή αντί της διαμέσου.
2. Η διάμεσος και το δεύτερο τεταρτημόριο έχουν πάντα την ίδια τιμή.

7) Ερώτηση:
Ποια είναι η διαφορά μεταξύ διακύμανσης και της τυπικής απόκλισης;

8) Ερώτηση:
Στον παρακάτω πίνακα έχουμε τα αποτελέσματα 3 δημόσιων υπαλλήλων σε τέσσερα τεστ δεξιοτήτων, σχετικά με τη γλωσσική, μαθηματική, καλλιτεχνική και τεχνολογική δεξιότητά τους.

Δημόσιοι Υπάλληλοι	Γλωσσική Δεξιότητα	Μαθηματική Δεξιότητα	Καλλιτεχνική Δεξιότητα	Τεχνολογική Δεξιότητα	Μέσος Όρος
1	8	10	8	10	9
2	9	9	9	9	9
3	8	9	9	10	9

Παρατηρούμε ότι και οι τρεις δημόσιοι υπάλληλοι αν και δεν είχαν την ίδια βαθμολογία είχαν το ίδιο μέσο όρο. Υπολογίστε τη διακύμανση του κάθε δημοσίου υπαλλήλου. Σε τι συμπέρασμα καταλήγετε;

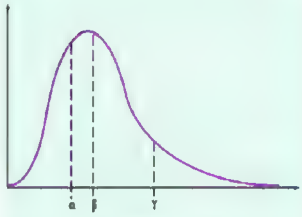
9) Ερώτηση:
Να χαρακτηρίσετε τις προτάσεις, που ακολουθούν με (Σ), στην περίπτωση, που θεωρείται, πως είναι Σωστές και (Λ), στην αντίθετη περίπτωση.

Α. Το εύρος είναι μέτρο θέσης.
Β. Η διακύμανση εκφράζεται με τις ίδιες μονάδες με τις οποίες εκφράζονται οι παρατηρήσεις.
Γ. Το κυκλικό διάγραμμα χρησιμοποιείται μόνο για τη γραφική παράσταση των ποσοτικών μεταβλητών.

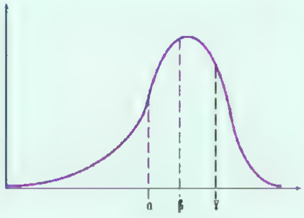
10) Ερώτηση:
1) Έχουμε ένα δείγμα ν=10 παρατηρήσεων, όπου κάθε παρατήρηση μπορεί να είναι 1, 2 ή 3. Είναι δυνατό η μέση τιμή να είναι
α) 1 β) 4 γ)1,8

2) Αν για 20 μετρήσεις οι πιθανές τιμές τους είναι 0 ή 1, να εξεταστεί κατά πόσο, η μέση τιμή των μετρήσεων, μπορεί να είναι ίση με
α) 1,2 β) 0,9 γ) 1,5

11) Ερώτηση:
Να επιλεγεί, σε κάθε από τα παρακάτω διαγράμματα, ένα από τα σημεία α, β ή γ, έτσι ώστε αυτό να αντιπροσωπεύει την διάμεσο της κατανομής.



διάγραμμα 1



διάγραμμα 2

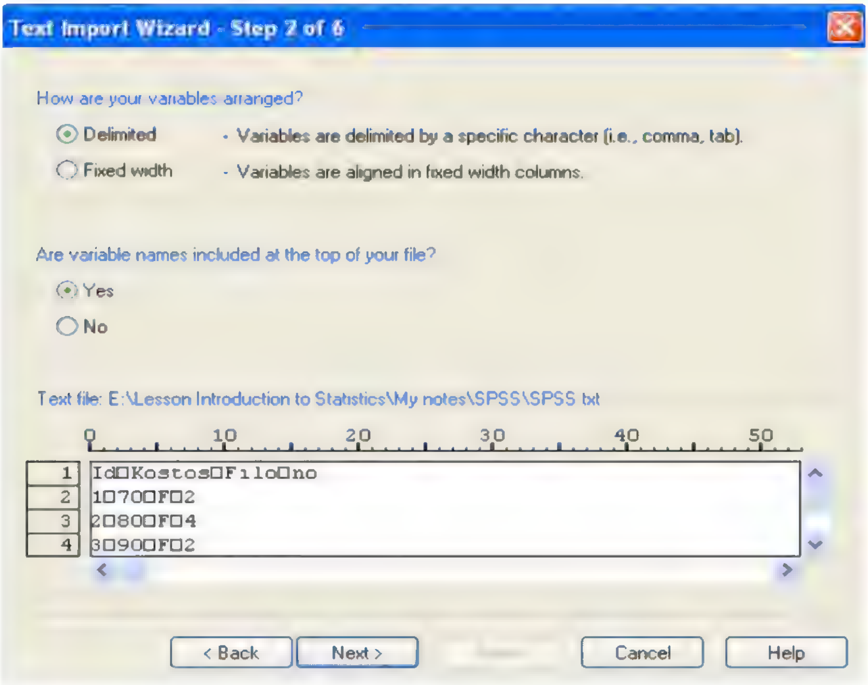
διάγραμμα 3

12) Ερώτηση:
Οι μέθοδοι ταυτόχρονης παρουσίασης .... τουλάχιστον ....χαρακτηριστικών (μεταβλητών) περιορίζονται στους πίνακες .... με τη διαδικασία "................. " του SPSS. Μπορούμε να χρησιμοποιήσουμε τη διαδικασία "................... " η οποία θα μας δώσει επιπλέον μια εικόνα του συνόλου των παρατηρήσεων. Με τη διαδικασία " ...." μπορούμε να πετύχουμε την πιο πλούσια και πλήρη περιγραφική στατιστική των παρατηρήσεων μιας .... μεταβλητής μέσα στις διάφορες κατηγορίες κάποιας ποιοτικής.

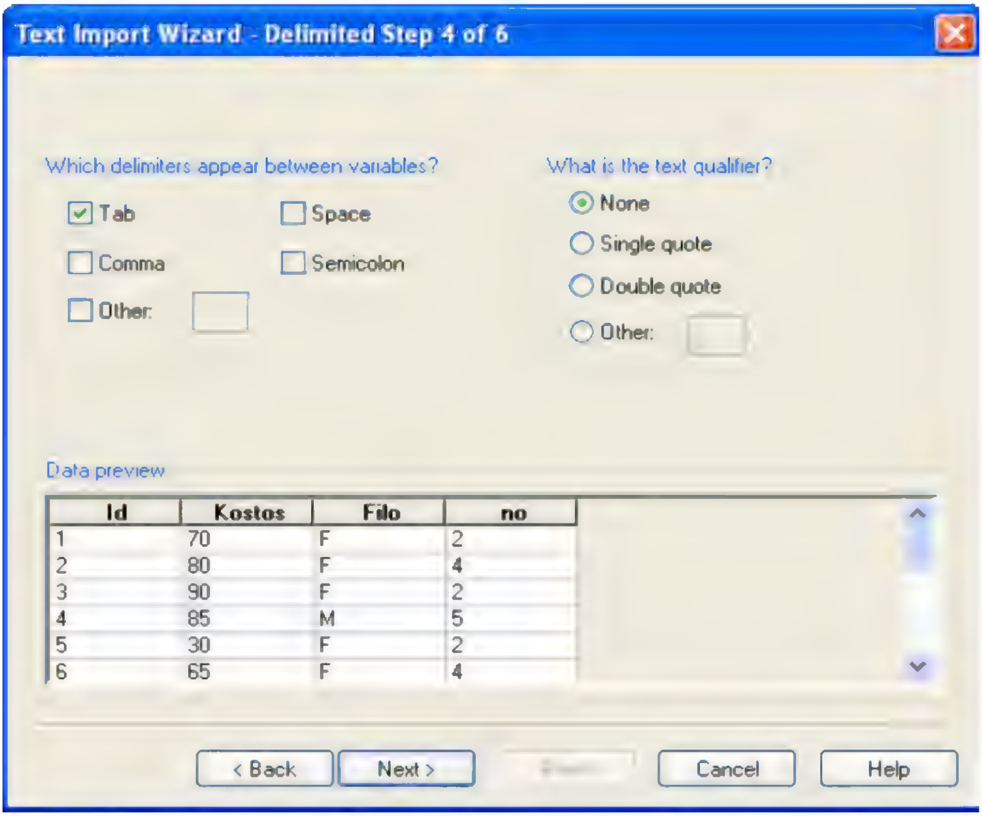
13) Ερώτηση:
Όταν μια κατανομή είναι απόλυτα ...., η μέση τιμή και η διάμεσος .... Όταν η κατανομή έχει .... ασυμμετρία η μέση τιμή είναι δεξιότερα από τη διάμεσο, ενώ όταν η κατανομή έχει .... στρεβλότητα η μέση τιμή είναι αριστερότερα από τη διάμεσο.

14) Ερώτηση:
Να αναφέρετε αναλυτικά τα βήματα που θα πρέπει να ακολουθήσετε για να προκύψει πίνακας με συγκεκριμένη μορφή και συγκεκριμένα μέτρα θέσης – διασποράς. (Μπορείτε να δημιουργήσετε ένα πίνακα της επιλογής σας).

15) Ερώτηση:
Στην περίπτωση που έχουμε να εισαγάγουμε στο SPSS αρχείο ASCHII (κείμενο), να περιγράψετε η κάθε εικόνα τι δηλώνει πως θα εφαρμόσουμε κάθε φορά.



Εικόνα Α
Η εικόνα Α δηλώνει...............



Εικόνα Β

Η εικόνα Β δηλώνει...............

16) Ερώτηση:

Πώς διαφοροποιούνται τα αρχεία ASCHII κατά την εισαγωγή τους σε βάση SPSS;

17) Ερώτηση:

Ο δειγματικός μέσος μπορεί να θεωρηθεί ως κατάλληλη εκτιμήτρια.

Σωστό

Λάθος

18) Ερώτηση:
Μια εκτιμήτρια $\hat{\theta}_n$ μπορεί να θεωρηθεί ως συνεπής εκτιμήτρια μιας παραμέτρου θ του πληθυσμού αν για μια οποιαδήποτε πολύ μικρή θετική τιμή ε ισχύει ότι:

$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$ ανεξάρτητα από το μέγεθος του δείγματος n.

Σωστό
Λάθος

19) Ερώτηση:
Η μέθοδος των ροπών ουσιαστικά βασίζεται στον υπολογισμό τόσων ροπών όσες οι εκτιμώμενες παράμετροι.
Σωστό
Λάθος

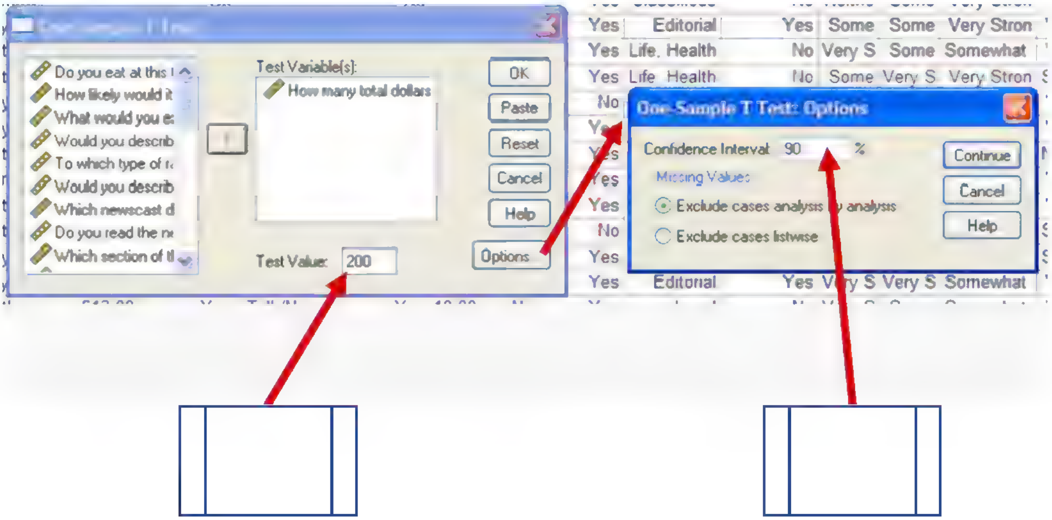
20) Ερώτηση:
Δώστε τον ορισμό της ελεγχοσυνάρτησης.

21) Ερώτηση:
Στην περίπτωση κατασκευής διαστημάτων εμπιστοσύνης για τη μέση τιμή ενός κανονικού πληθυσμού χρειάζεται να γνωρίζουμε τη διακύμανση του πληθυσμού.
Σωστό
Λάθος

22) Ερώτηση:
Στην περίπτωση ελέγχου υποθέσεων για τη σύγκριση μέσων τιμών ενός κανονικού πληθυσμού μπορούμε να χρησιμοποιήσουμε το κεντρικό οριακό θεώρημα ώστε να προχωρήσουμε στον έλεγχο.
Σωστό
Λάθος

23) Ερώτηση:

Να συμπληρώσετε τα κενά πλαίσια, αποτυπώνοντας τι υπολογίζει, η συγκεκριμένη οθόνη του SPSS.



24) Ερώτηση:

Από τον παρακάτω πίνακα αποτελεσμάτων:

1. Να αποτυπώσετε την μηδενική και εναλλακτική υπόθεση.
2. Σε τι συμπέρασμα καταλήγετε από την τιμή του p-value.
3. Να υπολογίσετε το 90% Διάστημα Εμπιστοσύνης για τη μέση τιμή.

One-Sample Test

	Test Value = 200					
					90% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
How many total dollars do you spend	-10,775	399	,000	-$49.94750	-$57.5897	-$42.3053

25) Ερώτηση:
Η Ανάλυση Διακύμανσης κατά ένα παράγοντα εφαρμόζεται για τον έλεγχο στατιστικών υποθέσεων χρησιμοποιώντας ως στατιστικό μέτρο τις διακυμάνσεις.
Σωστό
Λάθος

26) Ερώτηση:
Συμπληρώστε το παρακάτω κείμενο.

Η ανάλυση διακύμανσης ανήκει στην κατηγορία των .... Σκοπός της είναι να εντοπίσει την κύρια πηγή..................... Η υπόλοιπη διακύμανση θα αποδοθεί σε..... .... και ονομάζεται................

27) Ερώτηση
Η συνολική τετραγωνική απόκλιση εντός των δειγμάτων είναι το άθροισμα των τετραγώνων των αποκλίσεων των παρατηρήσεων από τη μέση τιμή.
Σωστό
Λάθος

28) Ερώτηση:
Η συνολική τετραγωνική απόκλιση μεταξύ των δειγμάτων υπολογίζεται ως:

$$\sum_i \sum_j Y_{ij}^2 - \sum_i \frac{Y_{i.}^2}{n_i}.$$

Σωστό
Λάθος

29) Ερώτηση:
Στην περίπτωση του πλήρως τυχαιοποιημένου πληθυσμού δε χρειάζεται καμία προϋπόθεση προκειμένου να εφαρμόσουμε τη μέθοδο της ανάλυσης διακύμανσης.
Σωστό
Λάθος

30) Ερώτηση:
Συμπληρώστε το παρακάτω κείμενο.

Ο όρος συνάφεια αναφέρεται στον τρόπο που δύο ή περισσότερες μεταβλητές.......................... και καθορίζει την....................... και το......................... που παρατηρείται στις τιμές μιας μεταβλητής όταν παρουσιάζονται αλλαγές στις τιμές μιας άλλης μεταβλητής.

31) Ερώτηση:
Απόλυτη συνάφεια ονομάζεται η συνάφεια που διατηρεί την ίδια κατεύθυνση σε όλο το εύρος της κλίμακας μέτρησης των μεταβλητών.
Σωστό
Λάθος

32) Ερώτηση:
Με τον έλεγχο ανεξαρτησίας μεταβλητών ουσιαστικά ελέγχουμε κατά πόσο το ενδεχόμενο μια παρατήρηση που ανήκει στην i γραμμή (μεταβλητή Α) να είναι ανεξάρτητο από το ενδεχόμενο η ίδια παρατήρηση να ανήκει ταυτόχρονα και στη j στήλη (μεταβλητή Β).
Σωστό
Λάθος

33) Ερώτηση:
Στους μη παραμετρικούς ελέγχους δε χρειάζεται να γνωρίζουμε την κατανομή των πληθυσμών από την οποία έχουν προέλθει τα δείγματα μας.
Σωστό
Λάθος

34) Ερώτηση:
Ποιος από τους παρακάτω ελέγχους είναι μη παραμετρικός;
i) Chi - Square
ii) Kolmogorov - Smirnov
iii) Runs Test
iv) Όλοι οι παραπάνω

35) Ερώτηση:
Ο Runs Test βασίζεται στον αριθμό των ροών που εμφανίζονται σε κάποιο δείγμα παρατηρήσεων.
Σωστό
Λάθος

36) Ερώτηση:
Ο Kolmogorov – Smirnov δεν επηρεάζεται από τις παραμέτρους της κατανομής που θέλουμε να ελέγξουμε.
Σωστό
Λάθος

37) Ερώτηση:
Με το διάγραμμα διασποράς μπορούμε να εντοπίσουμε την ύπαρξη συσχέτισης μεταξύ δύο μεταβλητών.
Σωστό
Λάθος

38) Ερώτηση:
Συμπληρώστε το παρακάτω κείμενο.

Ο όρος...................... αναφέρεται στον τρόπο που αλληλεπιδρά μια μεταβλητή με μία άλλη. Δύο είναι οι βασικές κατηγορίες.................... Η.......................... και η............................ Στην πρώτη περίπτωση τα ζεύγη σημείων (x_i, y_i) τείνουν να συσσωρεύονται γύρω από..................., ενώ στη δεύτερη γύρω. από...........................

39) Ερώτηση:
Η μηδενική συσχέτιση ερμηνεύεται ως απουσία οποιασδήποτε συσχέτισης μεταξύ δύο μεταβλητών.
Σωστό
Λάθος

40) Ερώτηση:
Συμπληρώστε το παρακάτω κείμενο.

Ο δειγματικός συντελεστής συσχέτισης παίρνει τιμές στο διάστημα:...... ≤ r ≤........ . Όσο πλησιέστερα στο............ είναι το r τόσο πιο ισχυρή......................... συσχέτιση έχουμε μεταξύ των Χ και Υ. Αντίθετα, τιμές του r πλησιέστερα στο.......... υποδηλώνουν ισχυρή αρνητική γραμμική εξάρτηση. Όταν το r τείνει να παίρνει τιμές κοντά στο................ αυτό είναι ένδειξη ασθενούς γραμμικής σχέσης μεταξύ των μεταβλητών.

41) Ερώτηση:
Ποια από τις παρακάτω μεθόδους πολλαπλής παλινδρόμησης θεωρείται ως η πιο αποτελεσματική;

i) Forward Procedure
ii) Stepwise Regression
iii) Backward Elimination
iv) Καμία από τις παραπάνω

42) Ερώτηση:
Στην πολλαπλή παλινδρόμηση κάποια από τις ανεξάρτητες είναι δυνατόν να έχει προέρθει από το συνδυασμό των άλλων ανεξάρτητων μεταβλητών (π.χ γινόμενο δύο μεταβλητών).
Σωστό
Λάθος

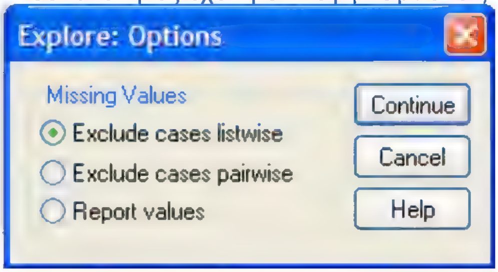
43) Ερώτηση:
Να τοποθετήσετε με τη σειρά τα βήματα του αλγόριθμου K-means.

Α) Βήμα: Κατάταξε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση
Β) Βήμα: Αν τα νέα κέντρα δεν διαφέρουν από τα παλιά σταμάτα αλλιώς πήγαινε στο βήμα 2.
Γ) Βήμα: Από τις παρατηρήσεις που είναι μέσα στην ομάδα υπολόγισε τα νέα κέντρα.
Δ) Βήμα: Βρες τα αρχικά κέντρα

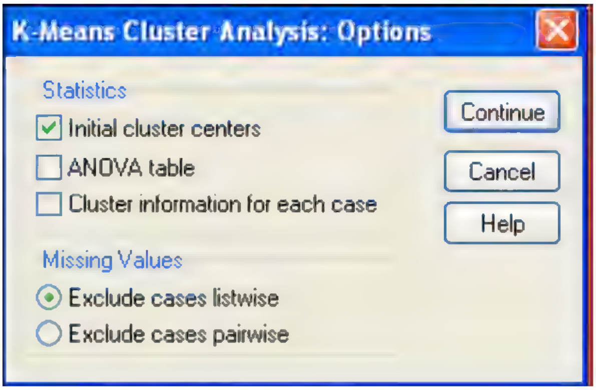
44) Ερώτηση:
Να τοποθετήσετε με τη σειρά τα βήματα του αλγόριθμου της ιεραρχικής ομαδοποίησης.

Α) Βήμα: Αν δεν έχουν όλες οι παρατηρήσεις μπει σε μια ομάδα πήγαινε στο βήμα 1 αλλιώς σταμάτα.
Β) Βήμα: Βρες τη μικρότερη απόσταση και ένωσε τις δύο παρατηρήσεις με τη μικρότερη απόσταση. Δηλαδή δημιουργούμε μια ομάδα με τις παρατηρήσεις που είναι πιο κοντά. Αν η μικρότερη απόσταση αφορά μια ήδη δημιουργηθείσα ομάδα και μια παρατήρηση απλά βάζουμε αυτή την παρατήρηση σε αυτή την ομάδα ή αν αφορά 2 ομάδες που ήδη υπάρχουν τις ενώνουμε.
Γ) Βήμα: Δημιούργησε τον πίνακα αποστάσεων για όλες τις ομάδες

45) Ερώτηση:
Τι δυνατότητες έχει η επιλογή "Options";



46) Ερώτηση:
Ποιες δυνατότητες, μας δίνει το παρακάτω πλαίσιο διαλόγου του SPSS;



ΑΠΑΝΤΗΣΕΙΣ

1) Απάντηση:
Πληθυσμός (population) είναι ένα σύνολο στοιχείων που μας ενδιαφέρει να μελετήσουμε. Πολλές φορές χρησιμοποιείται ο όρος **ολότητα (universe)**.
Δείγμα (sample) είναι υποσύνολο ενός πληθυσμού ή παρατηρηθέντων αποτελεσμάτων μίας διαδικασίας για μια χρονική περίοδο.

2) Απάντηση:
Η στατιστική σκέψη είναι μια διαδικασία συλλογισμών, που αναγνωρίζει ότι υπάρχει μεταβλητότητα σε όλα τα φαινόμενα και η μελέτη της μεταβλητότητας οδηγεί σε νέες γνώσεις και καλύτερες αποφάσεις.

3) Απάντηση:
α) Πιθανόν να έχουν τις ίδιες πολιτικές πεποιθήσεις.
β) Θα έχουμε υπερεκτίμηση του εισοδήματος.

4) Απάντηση:
(ε)

5) Απάντηση:

$$\bar{x} = 6 \Leftrightarrow \frac{7+5+\alpha+2+5+\beta+8+6+\gamma+5+3}{11} = 6 \Leftrightarrow \alpha+\beta+\gamma = 25$$

6) Απάντηση:
(Λ), (Σ)

7) Απάντηση:
Η διακύμανση είναι το τετράγωνο της τυπικής απόκλισης. Συνήθως χρησιμοποιούμε την τυπική απόκλιση διότι αυτή εκφράζεται στις μονάδες της μεταβλητής, ενώ η διακύμανση στα τετράγωνα των μονάδων αυτών και κατά συνέπεια δεν είναι ερμηνεύσιμη.

8) Απάντηση:
Για να απαντήσουμε στο παραπάνω ερώτημα αρκεί να υπολογίσουμε τις αντίστοιχες διακυμάνσεις.

Η διακύμανση του 1ου είναι :
$[(8-9)^2+(10-9)^2+(8-9)^2+(10-9)^2]/4 = (1+1+1+1)/4 = 1$

Του 2ου είναι:
$[(9-9)^2+(9-9)^2+(9-9)^2+(9-9)^2]/4 = 0/4 = 0$

και του 3ου είναι:
$[(8-9)^2+(9-9)^2+(9-9)^2+(10-9)^2]/4 = (1+0+0+1)/4 = 2/4 = 0.5$.

Συνεπώς μεγαλύτερη σταθερότητα παρουσιάζει ο 2ος ακολουθεί ο 3ος και 1ος.

9) Απάντηση:
(Α), (Β), (Γ) Λάθος

10) Απάντηση:
1) (α) και (γ)
2) (β)
Σημείωση: Η μέση τιμή είναι πάντα ανάμεσα στη μικρότερη και μεγαλύτερη παρατήρηση.

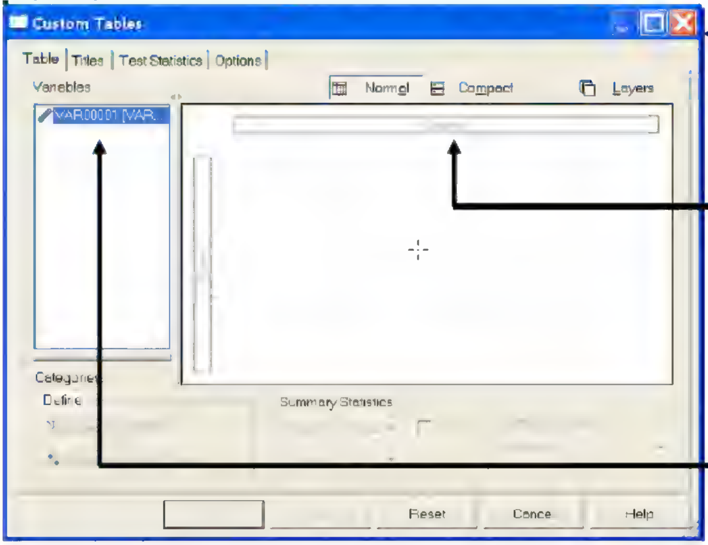
11) Απάντηση:
(α), (β), (β)

12) Απάντηση:
Δύο, Ποιοτικών, Συνάφειας, Crosstabs, Case Summaries, Explore, ποσοτικής.

13) Απάντηση:
Συμμετρική, συμπίπτουν, δεξιά, αριστερή.

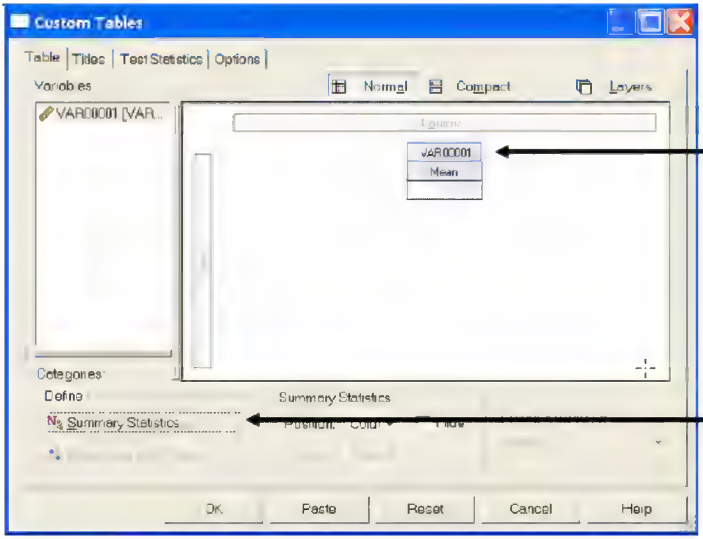
14) Απάντηση:

Αρχικά επιλέγουμε Analyze-Tables-Customs Tables και προκύπτουν τα ακόλουθα παράθυρα:



Θα ανοίξει το διπλανό παράθυρο.

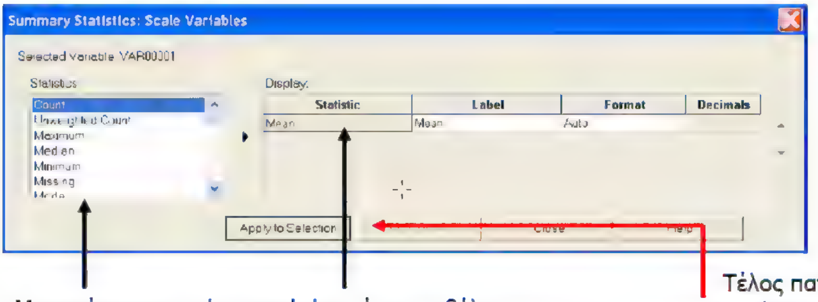
Με Drug-and-Drop μεταφέρουμε τη μεταβλητή από το αριστερό πλαίσιο στο δεξιό πλαίσιο στη θέση **columns**



Η μεταβλητή έχει μεταφερθεί κάτω από τον τίτλο columns

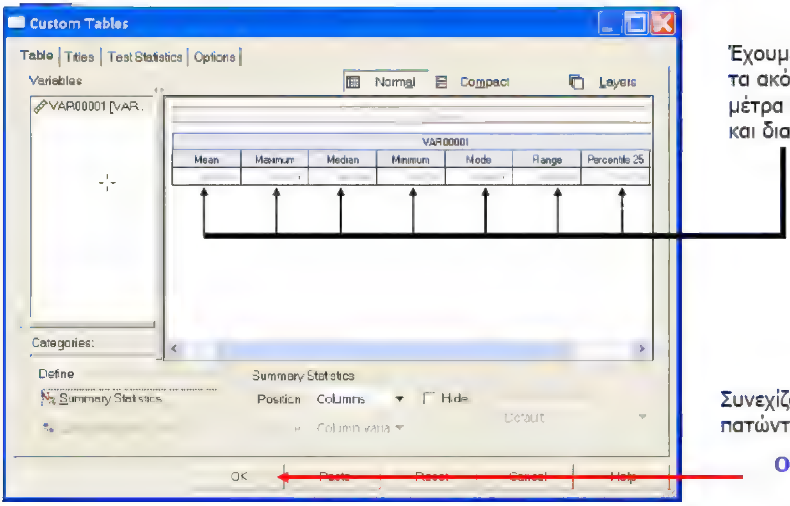
Επιλέγουμε:

Summary statistics...



Μεταφέρουμε με drug-and-drop ή με το βέλος τα περιγραφικά στατιστικά που θέλουμε από την αριστερή λίστα στο δεξί τμήμα

Τέλος πατάμε apply to selection και επιστρέφουμε στο προηγούμενο παράθυρο



Έχουμε επιλέξει τα ακόλουθα μέτρα θέσης και διασποράς:

Συνεχίζουμε πατώντας **OK….**

15) Απάντηση:
Εικόνα Α: Η μορφή των δεδομένων καθορίζεται ως Tab delimited (οι στήλες χωρίζονται από την παράγραφο) – τα ονόματα των στηλών βρίσκονται στην κορυφή του αρχείου.
Εικόνα Β: Επιβεβαιώνουμε πως τα αρχεία χωρίζονται με το tab και βλέπουμε πως τα δεδομένα θα εισαχθούν σε στήλες.

16) Απάντηση:
Διαφοροποιείται το δεύτερο βήμα ανάλογα με το αν έχουμε περίπτωση διαχωρισμένων μεταβλητών ή περίπτωση σταθερού εύρους μεταβλητών. Στην πρώτη περίπτωση επιλέγουμε στο βήμα 2 delimited ενώ στη δεύτερη επιλέγουμε Fixed with.

17) Απάντηση:
Σωστό

18) Απάντηση:
Λάθος

19) Απάντηση:
Σωστό

20) Απάντηση:
Ελεγχοσυνάρτηση ονομάζεται η τιμής της στατιστικής συνάρτησης που χρησιμοποιούμε για να μετρήσουμε τη διαφορά των δεδομένων από αυτό που αναμένουμε να συμβεί όταν η μηδενική υπόθεση είναι ακριβής.

21) Απάντηση:
Λάθος

22) Απάντηση:
Σωστό

23) Απάντηση:
Μηδενική Υπόθεση H_0:μ=200 και εναλλακτική H_1:μ≠200.
Και στο δεύτερο πλαίσιο απεικονίζει ότι το Διάστημα εμπιστοσύνης για τον παραπάνω έλεγχο είναι 90%.

24) Απάντηση:
1. Μηδενική Υπόθεση H_0:μ=200 και εναλλακτική H_1:μ≠200.
2. Απορρίπτουμε τη μηδενική υπόθεση, γιατί p-value<0,005.
3. Το 90% Διάστημα Εμπιστοσύνης είναι -57,5897 ≤ μ-200 ≤ -42,3053. Άρα 142,4103 ≤ μ ≤ 157,6947.

25) Απάντηση:
Λάθος

26) Απάντηση:
(πειραματικών σχεδιασμών), (διακύμανσης), (τυχαίους παράγοντες), (σφάλμα).

27) Απάντηση:
Σωστό

28) Απάντηση:
Λάθος

29) Απάντηση:
Λάθος

30) Απάντηση:
(αλληλεπιδρούν), (κατεύθυνση), (βαθμό μεταβολής).

31) Απάντηση:
Λάθος

32) Απάντηση:
Σωστό

33) Απάντηση:
Σωστό

34) Απάντηση:
iv)

35) Απάντηση:
Σωστό

36) Απάντηση:
Λάθος

37) Απάντηση:
Σωστό

38) Απάντηση:
(συσχέτιση), (συσχέτισης), (γραμμική), (μη γραμμική), (μια ευθεία γραμμή), (μια καμπύλη).

39) Απάντηση:
Λάθος

40) Απάντηση:
($-1 \leq r \leq 1$), (1), (θετική γραμμική), (-1), (μηδέν).

41) Απάντηση:
(ii)

42) Απάντηση:
Σωστό

43) Απάντηση:
Δ-Α-Γ-Β

44) Απάντηση:
Γ – Β – Α

45) Απάντηση:
Από την επιλογή "Options" μπορούμε να χειριστούμε τις ακραίες ή παράτυπες τιμές στο σύνολο των δεδομένων, που επεξεργαζόμαστε.

46) Απάντηση:
Initial Cluster Centers: Περιέχει τα αρχικά κέντρα των ομάδων, αυτά δηλαδή από όπου ξεκινά ο αλγόριθμος.

Iteration History: Περιέχει πληροφορίες για το πως μετακινείται ο αλγόριθμος σε κάθε επανάληψη. Η τιμή που εμφανίζεται είναι η απόσταση ανάμεσα στο κέντρο της ομάδας στην τρέχουσα επανάληψη με το κέντρο της ομάδας κατά την προηγούμενη. Όταν η απόσταση αυτή μηδενιστεί σταματά ο αλγόριθμος.

Final Cluster Centers: Περιέχει τα κέντρα των ομάδων που βρέθηκαν αφού σταμάτησε ο αλγόριθμος.

ANOVA: Ο πίνακας περιέχει την ανάλυση διακύμανσης για το αν διαφέρουν οι μέσες τιμές ανάμεσα στις ομάδες. Μεταβλητές με καλή ικανότητα να ξεχωρίζουν τις παρατηρήσεις πρέπει να είναι στατιστικά σημαντικές. Πρέπει να ληφθεί υπόψη πως αυτές οι τιμές της στατιστικής σημαντικότητας έχουν μάλλον περιγραφικό σκοπό για να συγκρίνουμε μεταβλητές μεταξύ τους καθώς ο αλγόριθμος έχει κατάλληλα σχεδιαστεί να μεγιστοποιεί την ελεγχοσυνάρτηση F και επομένως η χρήση του είναι μάλλον ενδεικτική.

Number of Cases in each Cluster: Ο πίνακας παρουσιάζει πόσες παρατηρήσεις περιέχει κάθε ομάδα τελικά.

ΠΑΡΑΡΤΗΜΑ I

1. Εφαρμογές με τη χρήση του στατιστικού πακέτου SPSS

1.1 Χρήση Διαγραμμάτων για την Αναπαράσταση Πληροφορίας στο SPSS

Ένα από τα πιο χρήσιμα στοιχεία που μπορεί να μας προσφέρει ένα λογισμικό στατιστικής ανάλυσης όπως το SPSS είναι τα γραφήματα. Με το SPSS μπορούμε να κατασκευάσουμε ραβδογράμματα, κυκλικά διαγράμματα (πίττες), ιστογράμματα, διαγράμματα διασποράς κ.λπ. Ανεξάρτητα από τον τύπο του γραφήματος που θέλουμε να έχουμε για τα δεδομένα μας, οι διαδικασίες δημιουργίας, επεξεργασίας, τροποποίησης, αποθήκευσης και εκτύπωσης ενός γραφήματος είναι περίπου οι ίδιες:

◇ **Δημιουργία γραφήματος.** Μπορούμε να δημιουργήσουμε ένα γράφημα με τη βοήθεια της επιλογής Graphs στη βασική ράβδο προτιμήσεων του λογισμικού (εικόνα 1.7.1.3). Επιπλέον, υπάρχουν κι αρκετές στατιστικές επεξεργασίες που παράγουν γραφήματα.

◇ **Εμφάνιση γραφήματος.** Οι επικεφαλίδες των γραφικών παραστάσεων που κατασκευάζουμε για το αρχείο δεδομένων που έχουμε, εμφανίζονται στο αριστερό τμήμα (display pane) του παραθύρου αποτελεσμάτων (Output Navigator) οπότε μπορούμε εύκολα να τις αναζητήσουμε για περαιτέρω επεξεργασία.

◇ **Επεξεργασία-τροποποίηση γραφήματος.** Για να επεξεργαστούμε ένα γράφημα, το επιλέγουμε και στη συνέχεια με διπλό κλικ ειδοποιούμε το λογισμικό ότι θέλουμε να επέμβουμε στην εμφάνισή του (με το βέλος του ποντικιού να βρίσκεται μέσα σε οποιοδήποτε σημείο του γραφήματος). Η γραφική παράσταση μεταφέρεται τότε σ' ένα ξεχωριστό παράθυρο, τον επεξεργαστή γραφημάτων του λογισμικού (Chart Editor), με τη βοήθεια του οποίου προχωρούμε στην όποια αλλαγή θέλουμε, όπως παρουσιάζεται παρακάτω.

◇ **Αποθήκευση του γραφήματος.** Τα γραφήματα αποθηκεύονται σαν τμήμα του αρχείου αποτελεσμάτων της στατιστικής ανάλυσης που πραγματοποιούμε. Με την τακτική Copy-Paste μπορούμε φυσικά να το αντιγράψουμε αρχικά στη μνήμη της μηχανής και στη συνέχεια όπου θελήσουμε.

◇ **Εκτύπωση γραφήματος.** Αφού επιλέξουμε τη γραφική παράσταση που μας ενδιαφέρει, από τη βασική ράβδο προτιμήσεων δίνουμε διαδοχικά:

File => Print => • Selection

Τα βασικά βήματα για τη δημιουργία ενός γραφήματος είναι όμοια και ανεξάρτητα από το είδος του. Αυτό που πρέπει να κατανοήσουμε είναι η σχέση μεταξύ της δομής του όποιου γραφήματος και της δομής των δεδομένων. Πιο κάτω παρουσιάζονται περιληπτικά όλοι σχεδόν οι τύποι γραφικών παραστάσεων. Τα δεδομένα που χρησιμοποιούνται στα διάφορα παραδείγματα προέρχονται από τα αρχεία της βάσης του στατιστικού πακέτου SPSS.

Ραβδογράμματα (Bar charts)

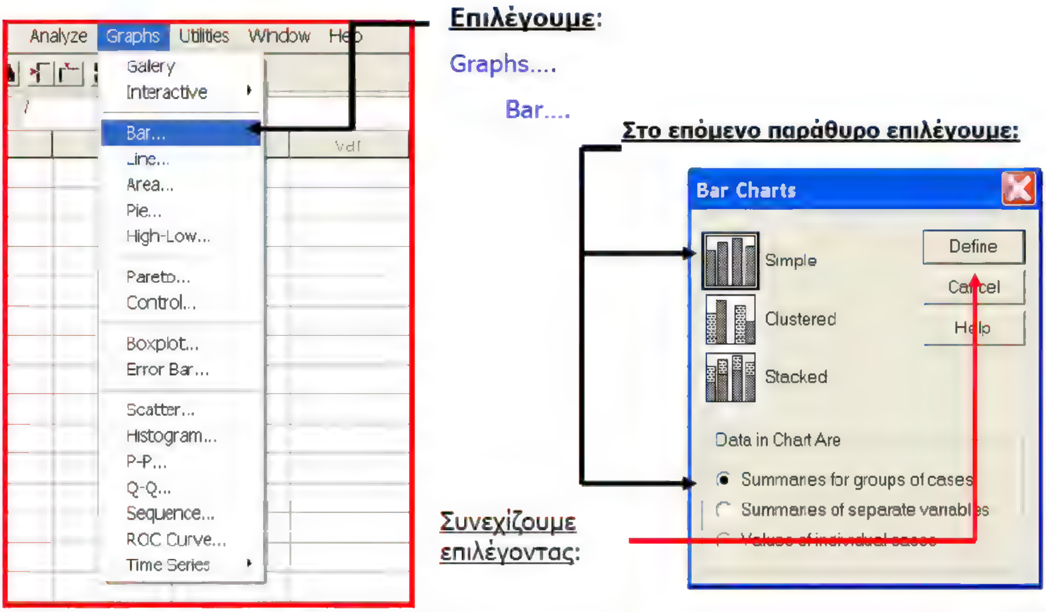
Τα **ραβδογράμματα (bar charts)** περιέχουν έναν εναλλακτικό τρόπο γραφικής παρουσίασης ποιοτικών δεδομένων. Το ραβδόγραμμα παριστά τη συχνότητα (ή τη σχετική συχνότητα) σε κάθε κατηγορία σαν ένα ορθογώνιο που είναι κάθετο στον

οριζόντιο άξονα. Το ύψος κάθε ορθογωνίου (ράβδου) είναι ανάλογο της συχνότητας (ή της σχετικής συχνότητας) της αντίστοιχης κατηγορίας. Δοθέντος ότι τα ραβδογράμματα αντιστοιχούν σε κατηγορίες ή σημεία παρά σε κλάσεις διαστημάτων (όπως τα ορθογώνια ενός ιστογράμματος που θα δούμε παρακάτω), το εύρος των ορθογωνίων μπορεί να είναι το ίδιο για όλα τα ορθογώνια (για όλες τις συχνότητες). Σε μερικές περιπτώσεις, προκειμένου το διάγραμμα να είναι περισσότερο σαφές, αφήνουμε χώρο μεταξύ των ράβδων (ορθογωνίων).

Σημαντική διαφορά των ραβδογραμμάτων από τα ιστογράμματα είναι ότι δεν υπάρχει φυσική σειρά με την οποία θα πρέπει να τοποθετηθούν οι κατηγορίες των ποιοτικών δεδομένων. Παρόλα αυτά, για προφανείς λόγους, είναι συχνό φαινόμενο να διατάσσουμε τις κατηγορίες ανάλογα με τις συχνότητες.

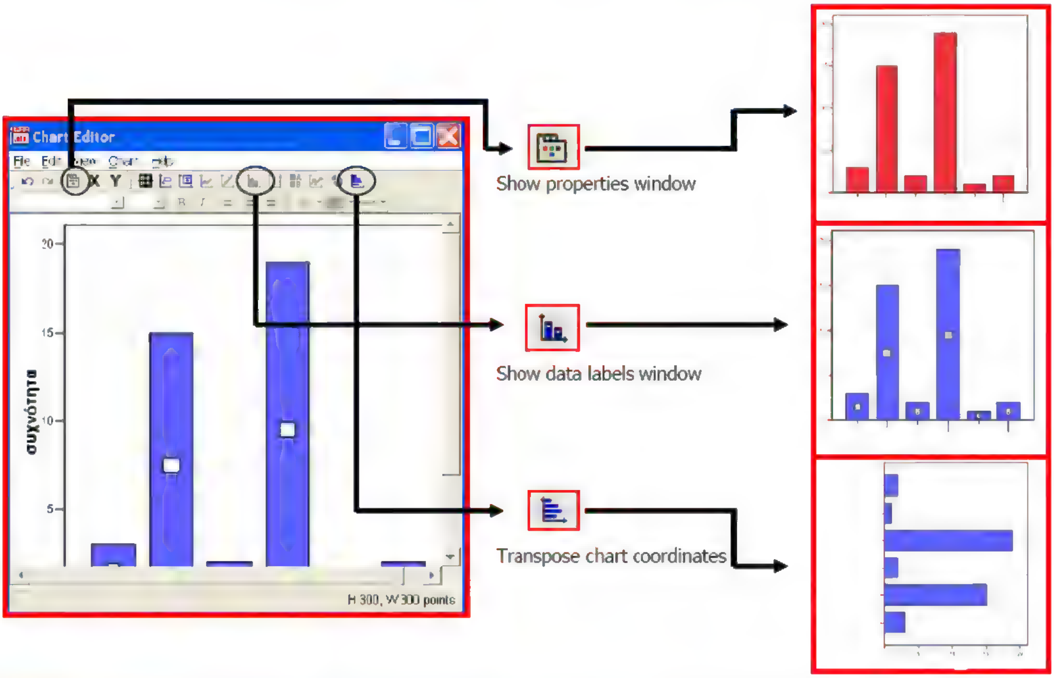
Το ραβδόγραμμα το οποίο απεικονίζει την κατανομή συχνότητας ποσοτικών δεδομένων με τις κατηγορίες διατεταγμένες ανάλογα με τη συχνότητα ονομάζεται **διάγραμμα Pareto (Pareto diagram)**.

Από τη βασική ράβδο προτιμήσεων του λογισμικού (εικόνα 1.1.1) επιλέγοντας διαδοχικά **Graphs => Bar** κατασκευάζουμε απλά (simple), ομαδοποιημένα (clustered) και συσσωρευμένα (stacked) ραβδογράμματα τα οποία συγκρίνουν την ίδια μεταβλητή μεταξύ κάποιων ομάδων περιπτώσεων (summaries for groups of cases) ή **μεταβλητές μεταξύ τους (summaries of separate variables)** ή τέλος περιπτώσεις μεταξύ τους (values of individual cases) (εικόνα 2.4.1). Η σύγκριση αυτή αναφέρεται συνήθως σε κάποιο στατιστικό μέτρο (εξ᾽ ορισμού τη μέση τιμή) κι επομένως σε ένα ραβδόγραμμα το μήκος της ράβδου παριστά την τιμή κάποιου στατιστικού ή μια ατομική τιμή. Επίσης, πατώντας διπλό κλικ πάνω στο γράφημα εμφανίζεται το παράθυρο Chart Editor στο οποίο μπορούμε να τροποποιήσουμε την εμφάνιση του γραφήματος (εικόνα 1.1.2).

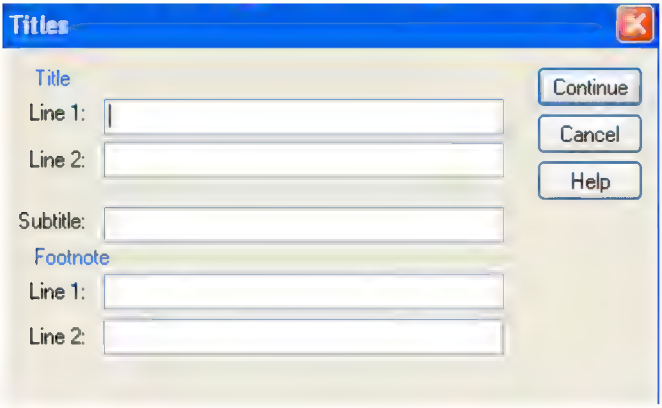


Εικόνα 1.1.1.: Βασική ράβδος διαδικασιών για την κατασκευή Ραβδογραμμάτων

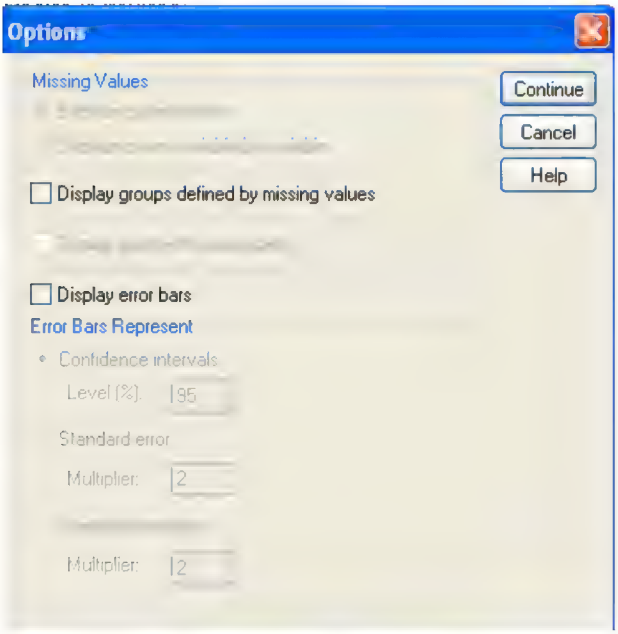
Η επιλογή **"Titles"** μας δίνει τη δυνατότητα να τοποθετήσουμε τους τίτλους στο διάγραμμα και η επιλογή "Options" μας επιτρέπει να εμφανίσουμε αν θέλουμε και ένα ακόμα ραβδόγραμμα που θα περιέχει το πλήθος των χαμένων τιμών.



Εικόνα 1.1.2.: Βοηθητικές επιλογές για την τροποποίηση των γραφημάτων (Chart Editor)



Εικόνα 1.1.3: Διαδικασία "Titles"



Εικόνα 1.1.4: Διαδικασία "Options"

Ιστογράμματα

Ένας εναλλακτικός τρόπος, ίσως περισσότερο γνωστός, παρουσίασης δεδομένων είναι η κατασκευή της κατανομής συχνότητας των δεδομένων και του αντίστοιχου ιστογράμματος. Για να κατασκευάσουμε την κατανομή συχνότητας διαιρούμε το διάστημα που καλύπτουν οι διαθέσιμες τιμές των δεδομένων σε μια σειρά από υποδιαστήματα (κατηγορίες, κλάσεις, τάξεις). Στη συνέχεια προσδιορίζουμε τον αριθμό των δεδομένων που περιέχονται σε κάθε κλάση. Σε αυτό το σημείο εισάγουμε την έννοια της συχνότητας (frequency), όπου είναι ο αριθμός των τιμών δεδομένων που βρίσκονται σε μια δεδομένη κλάση. Μερικές φορές είναι περισσότερο χρήσιμο να ξέρουμε το ποσοστό των τιμών των δεδομένων που βρίσκεται σε καθεμία κλάση αντί του καθεαυτού αριθμού. Σχετική συχνότητα (relative frequency) είναι το ποσοστό (proportion) όλων των τιμών των δεδομένων, που βρίσκονται σε μια δεδομένη κλάση. Η γραφική παράσταση που απεικονίζει τη συχνότητα ή τη σχετική συχνότητα σε σχέση με τα διαστήματα των κλάσεων ονομάζεται **ιστόγραμμα (histogram)**.

Τα βασικά σημεία τα οποία θα πρέπει να προσέχουμε όταν κατασκευάζουμε κατανομές συχνότητας είναι τα εξής:

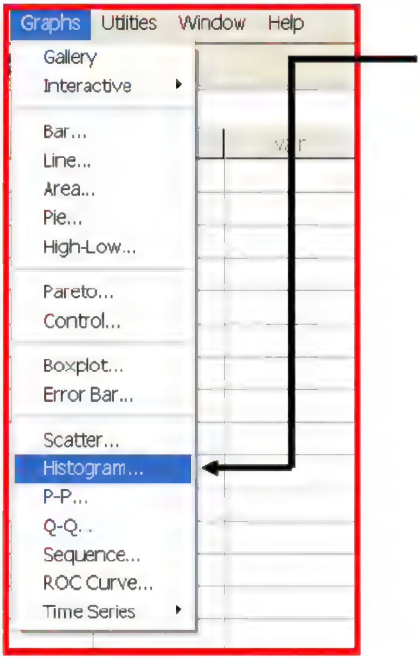
1....Οι κλάσεις που επιλέγουμε θα πρέπει να είναι τέτοιες που να δίνουν τη σωστή εικόνα της κατανομής των δεδομένων. Ο καθορισμός του αριθμού των κλάσεων είναι εν γένει αυθαίρετος. Συνήθως χρησιμοποιούμε από πέντε ως είκοσι κλάσεις. Όσο μεγαλύτερη ποσότητα δεδομένων είναι διαθέσιμη τόσο περισσότερες κλάσεις θα πρέπει να χρησιμοποιηθούν. Αυτό γιατί εάν ο αριθμός των κλάσεων που θα χρησιμοποιηθούν είναι πολύ μικρός ενδεχομένως να αποκρύπτουν σημαντικά χαρακτηριστικά των δεδομένων με την ομαδοποίησή τους. Από το άλλο μέρος, εάν ο αριθμός των κλάσεων είναι μεγάλος

σε σχέση με τα δεδομένα θα έχουμε πολλές κλάσεις που θα είναι ή κενές ή με μικρό αριθμό παρατηρήσεων και η κατανομή που θα εμφανίζουν θα δίνει μια όχι ικανοποιητική περιγραφή των δεδομένων.

2....Αφού δούμε το εύρος των τιμών του δείγματος θα πρέπει να καθορίσουμε το εύρος κάθε κλάσης (class width). Ως ένα γενικό κανόνα για το εύρος της κλάσης διαιρούμε τη διαφορά της μικρότερης από τη μεγαλύτερη μέτρηση με τον επιθυμητό αριθμό των κλάσεων που θέλουμε να χρησιμοποιήσουμε. Κατά κανόνα θα πρέπει να προσπαθούμε να έχουμε κλάσεις ίσου εύρους εκτός αν εμφανίζονται μετρήσεις με πολύ ακραίες τιμές, με αποτέλεσμα να χρησιμοποιούμε κλάσεις ανοικτού τύπου.

3....Καθορισμός των ορίων των κλάσεων: Είναι προφανές ότι τα όρια αυτά θα πρέπει να καθορίζονται με τρόπο ώστε οι μετρήσεις να κατανέμονται σε μία μόνο από τις δυνατές κατηγορίες.

Το ιστόγραμμα παριστάνει την κατανομή συχνοτήτων μιας μεταβλητής. Μοιάζει με το ραβδόγραμμα, αλλά εδώ οι ράβδοι είναι συνεχόμενες διότι χρησιμοποιείται για την παράσταση συχνοτήτων συνεχούς μεταβλητής. Κάθε ράβδος παριστάνει τη συχνότητα εμφάνισης των τιμών που βρίσκονται μέσα στο διάστημα που είναι η βάση της ράβδου. Το πλάτος των ορθογωνίων-ράβδων του ιστογράμματος μπορεί να ρυθμιστεί μέσα από τον επεξεργαστή γραφημάτων. Φυσικά αυτό συμπαρασύρει και το ύψος (συχνότητα) των τιμών που βρίσκονται μέσα στο διάστημα-βάση. Ιστογράμματα μπορούν να κατασκευαστούν μέσα από τις διαδικασίες "Frequencies" και "Explore" αλλά και από τη βασική ράβδο προτιμήσεων αν επιλέξουμε **Graphs => Histogram** (εικόνα 1.1.5).

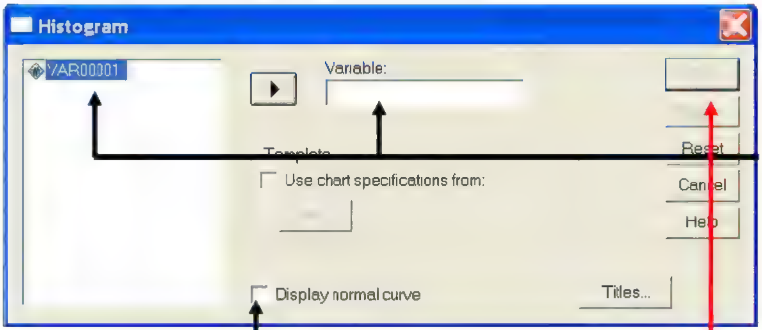


Επιλέγουμε:

Graphs....

Histogram....

Εικόνα 1.1.5: Βασική ράβδος διαδικασιών για την κατασκευή Ιστογραμμάτων.

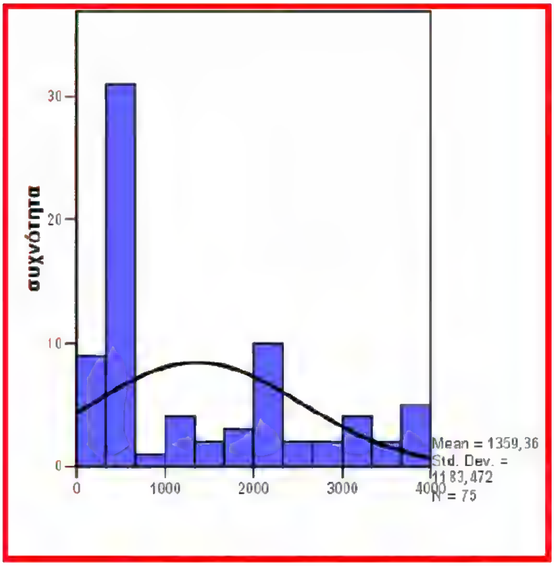


Μεταφέρουμε τη μεταβλητή επιλέγοντάς την και πατώντας το βέλος στο πλαίσιο: **Variable....**

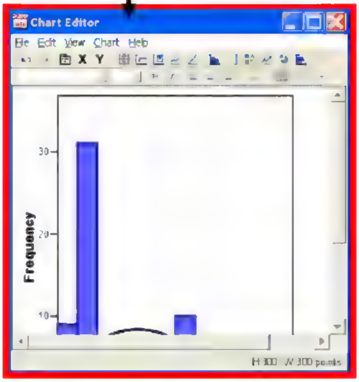
Τσεκάρουμε εάν θέλουμε παράλληλα με το ιστόγραμμα της κατανομής για λόγους σύγκρισης να εμφανίζεται και η συνάρτηση πυκνότητας πιθανότητας της **Κανονικής**

Συνέχεια πατώντας: **OK**

Εικόνα 1.1.6: Βασικές διαδικασίες στην κατασκευή ιστογραμμάτων.



Πατώντας διπλό αριστερό κλικ πάνω στο γράφημα εμφανίζεται το παράθυρο **Chart Editor** στο οποίο μπορούμε να τροποποιήσουμε την εμφάνιση του γραφήματος



Εικόνα 1.1.7: Chart Editor στην κατασκευή ιστογραμμάτων.

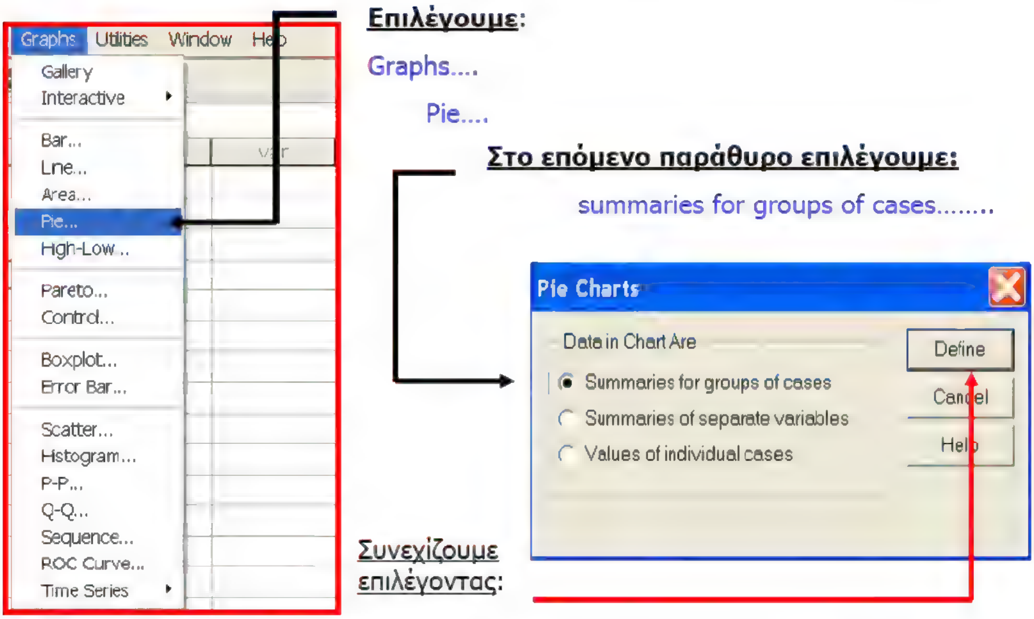
Κυκλικά διαγράμματα

Το **κυκλικό διάγραμμα (pie chart)** είναι απλά ένας κύκλος που υποδιαιρείται σε **κυκλικούς τομείς (slides)** οι οποίοι αντιπροσωπεύουν διάφορες κατηγορίες.

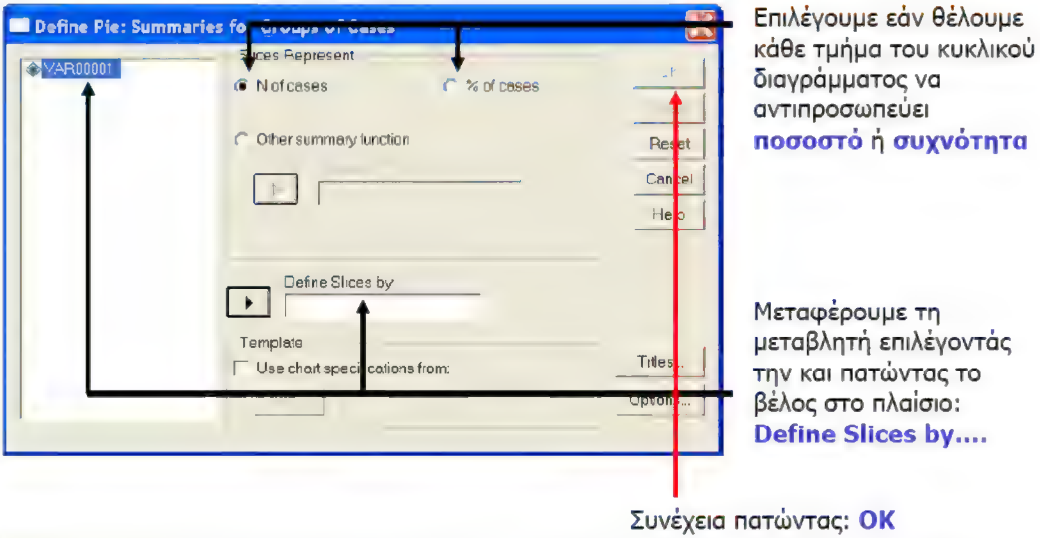
Τα κυκλικά διαγράμματα είναι αποτελεσματικά στις περιπτώσεις που αντικειμενικός σκοπός είναι να παρουσιαστούν οι συνιστώσες μίας ολότητας με τρόπο που αναδεικνύει τα σχετικά τους μεγέθη. Επειδή ακριβώς είναι πολύ παραστατικά, τα κυκλικά διαγράμματα χρησιμοποιούνται από εφημερίδες και περιοδικά για διάφορους λόγους και κυρίως όταν θέλουν να παρουσιάσουν οικονομικά μεγέθη. Τα κυκλικά διαγράμματα μπορούν να χρησιμοποιηθούν προκειμένου να συγκριθούν δύο διαφορετικές αναλύσεις ενός μεγέθους. Συχνά, για να είναι πιο εντυπωσιακή η παρουσίαση, οι κυκλικοί τομείς του διαγράμματος εμφανίζονται σε κάποια απόσταση από το κέντρο του κύκλου.

Τα κυκλικά διαγράμματα, γνωστά και σαν «πίττες», είναι ένας εύκολα κατανοητός τρόπος γραφικής παρουσίασης-σύγκρισης της ίδιας μεταβλητής μεταξύ κάποιων ομάδων περιπτώσεων, μεταβλητών μεταξύ τους ή τέλος περιπτώσεων μεταξύ τους. Η σύγκριση αυτή, γίνεται συγκρίνοντας το εμβαδόν κυκλικών τομέων ενός κύκλου.

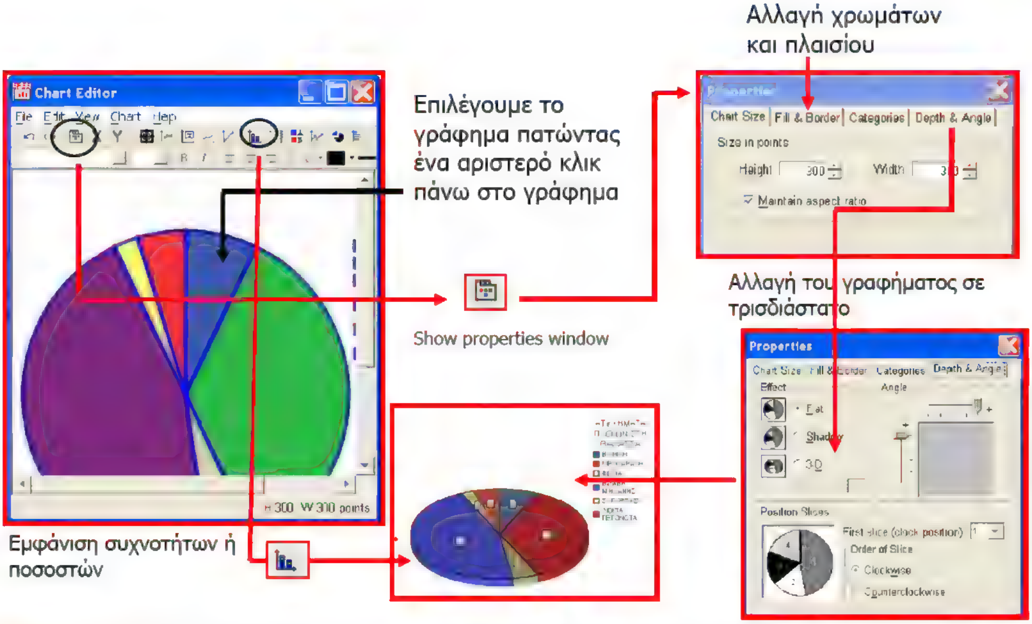
Για να κατασκευάσουμε ένα κυκλικό διάγραμμα επιλέγουμε από τη βασική ράβδο **Graphs Pie** (εικόνα 1.1.8). Υπάρχει η δυνατότητα να τονίσουμε κάποιον από αυτούς τους τομείς. Αρχικά μεταφέρουμε το γράφημα στον επεξεργαστή γραφημάτων (chart editor – εικόνα 1.1.10). Στη συνέχεια διαλέγουμε τον τομέα που μας ενδιαφέρει και από τη βασική ράβδο προτιμήσεων μορφοποιούμε το γράφημα σύμφωνα με τις προσωπικές μας επιλογές.



Εικόνα 1.1.8: Βασική ράβδος διαδικασιών για την κατασκευή των κυκλικών διαγραμμάτων



Εικόνα 1.1.9: Βασικές διαδικασίες στην κατασκευή κυκλικών διαγραμμάτων.



Εικόνα 1.1.10: Chart Editor στην κατασκευή κυκλικών διαγραμμάτων.

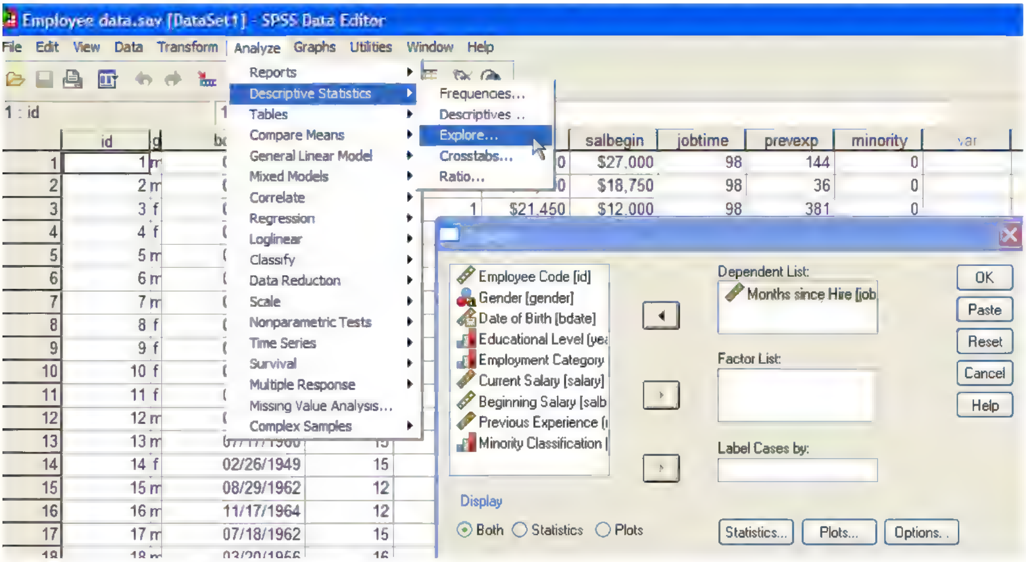
Φυλλογραφήματα

Τα διαγράμματα μίσχου-φύλλου ή φυλλογραφήματα (steam-and-leaf plots ή steam-and-leaf diagrams) είναι ένας πολύ απλός αλλά εξαιρετικά περιγραφικός τρόπος οργάνωσης και παρουσίασης δεδομένων με τρόπο που να περιγράφεται η κατανομή τους.

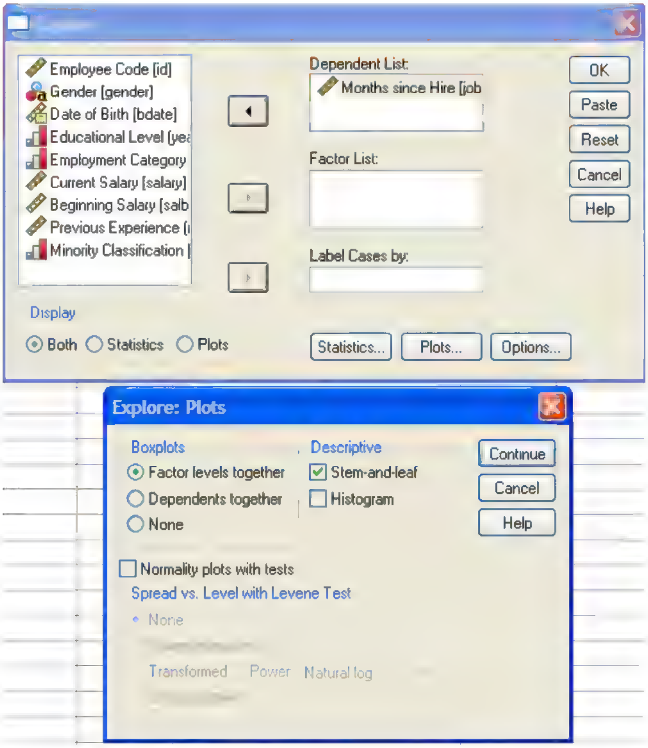
Η μέθοδος αυτή περιγραφής δεδομένων οφείλεται στο Στατιστικό John Tukey. Μπορεί να θεωρηθεί ως μια εναλλακτική μέθοδος για μια προκαταρκτική ανάλυση των στοιχείων της περιγραφής δεδομένων με ιστογράμματα. Μπορεί να αποτελεί ένα πρώτο βήμα στην κατασκευή μιας κατανομής συχνότητας και ενός ιστογράμματος.

Ο τρόπος παρουσίασης της κατανομής των δεδομένων με το συγκεκριμένο γράφημα είναι ο εξής: χωρίζουμε κάθε αριθμό από τα δεδομένα σε δύο μέρη: το μίσχο (stem) και το φύλλο (leaf). Στη συνέχεια κάνουμε μια κατακόρυφη γραμμή και αριστερά της τοποθετούμε το μίσχο (π.χ. το ψηφίο των δεκάδων) και στα δεξιά της το φύλλο. Ο καθορισμός του μίσχου και του φύλλου είναι αυθαίρετος. Η επιλογή γίνεται ώστε τα δεδομένα να παρουσιάζονται με τον καλύτερο τρόπο. Συνήθως τοποθετούμε τις τιμές των φύλλων σε κάθε γραμμή με τρόπο ώστε να αρχίζουν από τη μικρότερη προς τη μεγαλύτερη, για να είναι περισσότερο πληροφοριακό. Το μειονέκτημα είναι ότι υπάρχει η πιθανότητα να τοποθετηθούν λάθος οι αριθμοί. Υπάρχει και το ενδεχόμενο, για να είναι περισσότερο πληροφοριακό ένα γράφημα, να τοποθετούνται οι αριθμοί και από τις δύο πλευρές του κεντρικού μίσχου. Ο συγκεκριμένος τρόπος παρουσίασης ονομάζεται διάγραμμα μίσχου-φύλλου διπλής όψης (back to back stem-and-leaf plot). Για να κατασκευάσουμε ένα φυλλογράφημα ακολουθούμε τις παρακάτω επιλογές στο SPSS:

Analyze => Descriptive Statistics => Explore (εικόνα 1.1.11) και στη συνέχεια κάνουμε κλικ στην επιλογή "Plots" στο καινούριο παράθυρο, επιλέγουμε το διάγραμμα "Steam-and-leaf" (εικόνα 1.1.12).



Εικόνα 1.1.11: Βασική ράβδος διαδικασιών για την κατασκευή Φυλλογραφημάτων.



Εικόνα 1.1.12: Βασικές διαδικασίες στην κατασκευή των Φυλλογραφημάτων

Ας υποθέσουμε ότι θέλουμε να κατασκευάσουμε το Διάγραμμα Μίσχου-Φύλλου για το παράδειγμα με τις τελικές βαθμολογίες (με άριστα το 100) ενός τυχαίου δείγματος 15 δημόσιων υπαλλήλων που έχουν εγγραφεί στη θεματική ενότητα «Στατιστική Συμπερασματολογία με Στατιστικά Πακέτα» κατά το ακαδ. έτος 2007-08.
56, 67, 34, 89, 67, 89, 78, 57, 48, 47, 89, 80, 59, 89, 94

Χωρίζουμε κάθε παρατήρηση (βαθμολογία) σε δύο μέρη το μίσχο και το φύλλο. Στην προκειμένη περίπτωση μίσχος είναι το ψηφίο των δεκάδων και φύλλο το ψηφίο των μονάδων. Χαράσσουμε μια κατακόρυφη γραμμή και αριστερά της τοποθετούμε τους μίσχους (τα ψηφία των δεκάδων σε αύξουσα σειρά) δεξιά της τα αντίστοιχα φύλλα (τα αντίστοιχα ψηφία των μονάδων). Παρατηρούμε λοιπόν ότι:

- Η πρώτη τιμή του μίσχου είναι το 3 (η μικρότερη τιμή ψηφίου δεκάδων που συναντάται στο δείγμα των 15 βαθμολογιών) και η τιμή του φύλλου είναι 4 (το ψηφίο των μονάδων που αντιστοιχεί στο βαθμό 34).

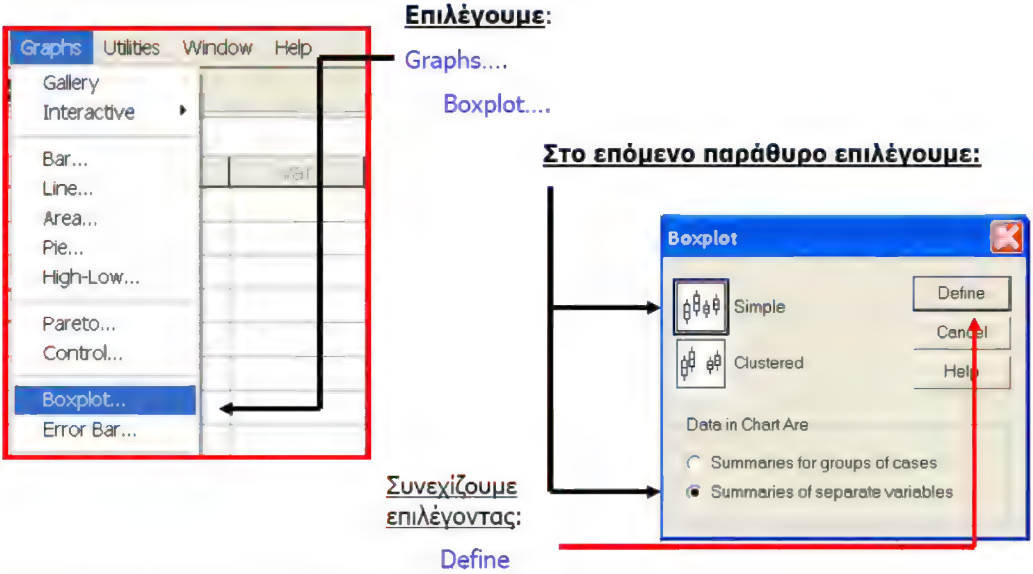
- Η επόμενη τιμή μίσχου είναι το 4 με τιμές φύλλων 7, 8 που αντιστοιχούν στο ψηφίο των μονάδων των βαθμών 47 και 48.

Η συνέχεια είναι προφανής και παρουσιάζεται στο διάγραμμα που ακολουθεί.

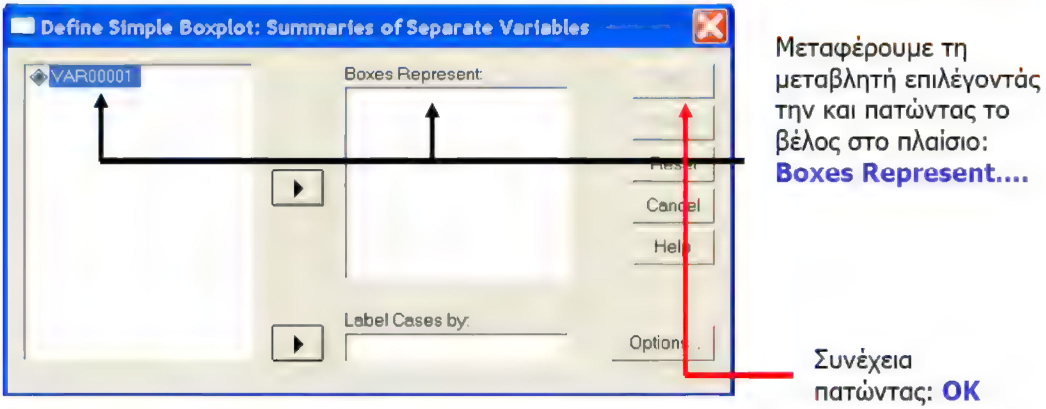
Μίσχος	Φύλλο
3	4
4	7 8
5	6 7 9
6	7 7
7	8
8	0 9 9 9 9
9	4

Θηκογράμματα

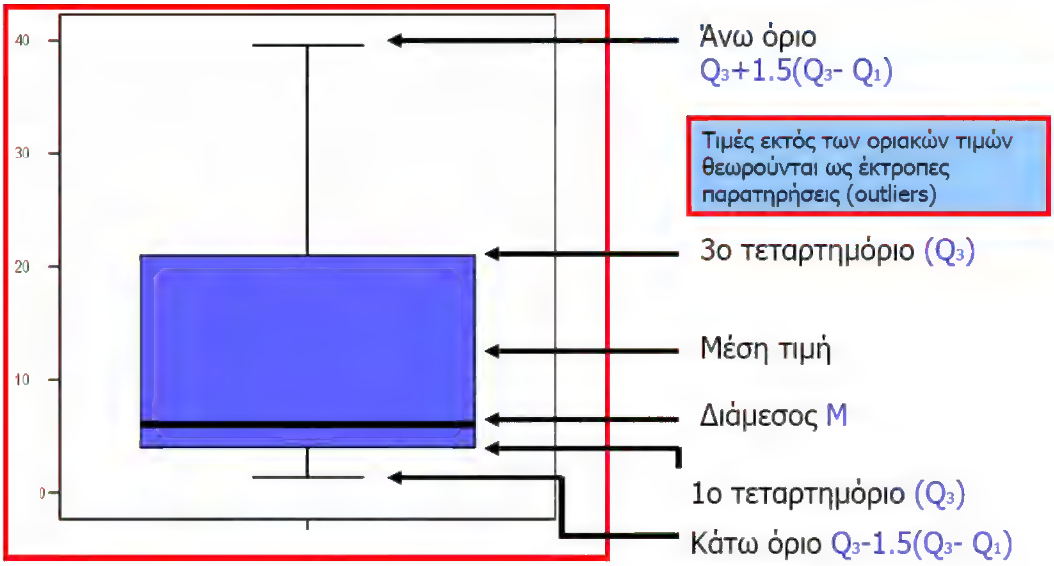
Τα θηκογράμματα χρησιμοποιούνται για την απεικόνιση της μεταβλητότητας κατανομών. Κάθε κατανομή αντιπροσωπεύεται με ένα ορθογώνιο παραλληλόγραμμο του οποίου το μήκος ισούται με το ενδοτεταρτημοριακό πλάτος των παρατηρούμενων τιμών. Στις βάσεις του ορθογωνίου τοποθετούνται δύο ευθύγραμμα τμήματα (ουρές) που εκτείνονται μέχρι τη μέγιστη και ελάχιστη παρατηρούμενη τιμή της κατανομής. Επιπλέον σημειώνεται και η διάμεσος (εικόνα 1.1.15). Οι παρατηρήσεις που βρίσκονται έξω από το ορθογώνιο χαρακτηρίζονται σαν παράτυπα σημεία (outliers) ή σαν ακραίες τιμές (extreme values). Το SPSS θεωρεί σαν παράτυπα σημεία τιμές που απέχουν από τις βάσεις του ορθογωνίου πάνω από 1.5 μήκη (του ορθογωνίου), και σαν ακραίες τιμές εκείνες που απέχουν πάνω από 3 μήκη. Με την επιλογή (εικόνα 1.1.13): **Graphs => Boxplot** κατασκευάζουμε απλά (simple) και ομαδοποιημένα (clustered) γραφήματα τα οποία συγκρίνουν την ίδια μεταβλητή μεταξύ κάποιων ομάδων περιπτώσεων (summaries for groups of cases) ή μεταβλητές μεταξύ τους (summaries of separate variables).



Εικόνα 1.1.13: Βασική ράβδος διαδικασιών για την κατασκευή Θηκογραμμάτων



Εικόνα 1.1.14: Βασικές διαδικασίες στην κατασκευή Θηκογραμμάτων



Εικόνα 1.1.15: Γραφική απεικόνιση Θηκογράμματος

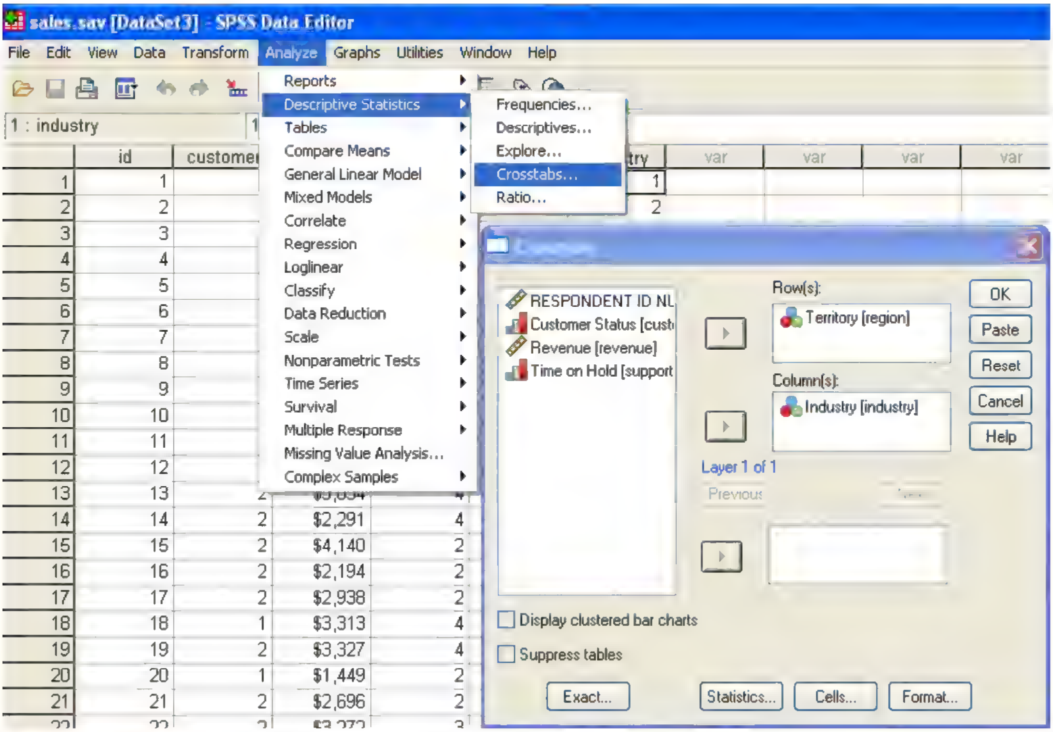
1.2 Παρουσίαση Πινάκων – Άναλυση των παρουσιάσεων μέσω των Συγκεντρωτικών πινάκων σε Ποιοτικά και Ποσοτικα χαρακτηριστικά.

Η ύπαρξη δύο ή περισσότερων μεταβλητών σε μια έρευνα, οδηγεί εύλογα στην αναζήτηση της (πιθανής) μεταξύ τους σχέσης. Η επιλογή της στατιστικής τεχνικής για τη συγκεκριμένη ανάλυση εξαρτάται αποκλειστικά από τη διάκριση των μεταβλητών σε ποιοτικές και ποσοτικές. Υπάρχη σαφής διαφοροποίηση των εργαλείων που είναι διαθέσιμα στην κάθε περίπτωση.

Άρχικα θα ασχοληθούμε με τα ποιοτικά χαρακτηριστικά. Οι μέθοδοι ταυτόχρονης παρουσιάσης δύο τουλάχιστον ποιοτικών χαρακτηριστικών (μεταβλητών) περιορίζονται στους πίνακες συνάφειας. Με τη διαδικασία **"Crosstabs"** του SPSS μπορούμε εκτός από την άμεση κατασκευή τους επιπλέον να προχωρήσουμε και στην αναζήτηση της έντασης και της φύσης της (πιθανής) σχέσης.

Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγοντας (εικόνα 1.2.1):

Analyze => Descriptive Statistics => Crosstabs



Εικόνα 1.2.1: Διαδικασία Crosstabs από τη βασική ράβδο

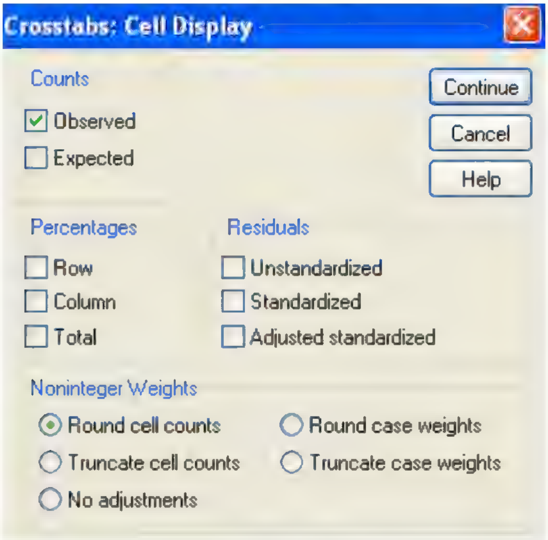
Διαλέγουμε την (ποιοτική) μεταβλητή, τις κατηγορίες της οποίας θέλουμε να έχουμε στις γραμμές του πίνακα συνάφειας και τη μετακινούμε στο παράθυρο Row(s). Διαλέγουμε κάποια άλλη (ποιοτική) μεταβλητή, τις κατηγορίες της οποίας θέλουμε να έχουμε στις στήλες του πίνακα συνάφειας και τη μετακινούμε στο παράθυρο Column(s).

Η παραπάνω διαδικασία μπορεί να χρησιμοποιηθεί και στην περίπτωση των ποσοτικών μεταβλητών των οποίων οι τιμές προηγουμένως έχουν κωδικοποιηθεί και αντιστοιχιστεί σε διαστήματα τιμών.

Φυσικά, μπορούμε να μετακινήσουμε περισσότερες από μια μεταβλητές τόσο στον κατάλογο Row(s) όσο και στον κατάλογο Column(s). Το SPSS θα κατασκευάσει από ένα διαφορετικό πίνακα συνάφειας για κάθε δυνατό συνδυασμό των μεταβλητών του καταλόγου Row(s) με εκείνες του καταλόγου Column(s).

Το SPSS παρέχει, με τη βοήθεια της επιλογής **"Layer"** τη δυνατότητα να ορίσουμε μία ή περισσότερες ποιοτικές μεταβλητές σαν μεταβλητές ελέγχου των ζητουμένων πινάκων συνάφειας: αρκεί να τη μετακινήσουμε στον κατάλογο "Layer". Το λογισμικό, για κάθε κατηγορία της πρώτης εκ των μεταβλητών αυτών θα κατασκευάσει όλους τους ζητούμενους στην προηγούμενη επιλογή πίνακες συνάφειας. Στη συνέχεια θα κάνει το ίδιο και για κάθε κατηγορία της δεύτερης ποιοτικής μεταβλητής που μετακινήσαμε στον κατάλογο "Layer" κ.ο.κ.

Αρχικά, επιλέγουμε τις πληροφορίες που θέλουμε να εμφανίζονται σε κάθε κελί του πίνακα συνάφειας (εικόνα 1.2.2). Εξ ορισμού, το λογισμικό θα εμφανίσει μόνον τις παρατηρούμενες συχνότητες. Με την επιλογή "Cells"' μπορούμε να εμπλουτίσουμε τις εμφανιζόμενες πληροφορίες με τις αναμενόμενες συχνότητες, τα ποσοστά των παρατηρήσεων στα κελιά της κάθε γραμμής ή στήλης, κ.λπ.



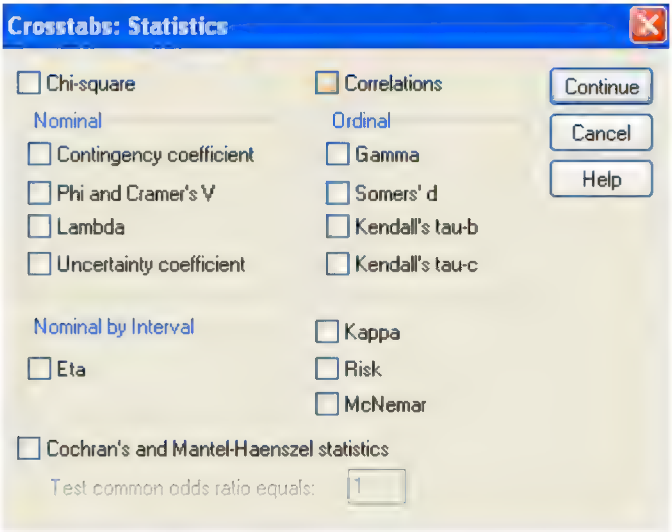
Εικόνα 1.2.2: Πληροφορίες που θέλουμε να είναι διαθέσιμες

√ **Counts:** Παρατηρούμενες συχνότητες (Observed) είναι οι μετρήσεις σε κάθε κελί, ενώ αναμενόμενες (Expected) είναι ο αριθμός των περιπτώσεων που θα ήταν σε κάθε κελί αν οι μεταβλητές που ορίζουν τις γραμμές και στήλες του πίνακα είναι στατιστικά ανεξάρτητες.

√ **Percentages:** Τα ποσοστά των γραμμών/στηλών αθροίζουν στο 100% κατά μήκος της κάθε γραμμής του πίνακα συνάφειας, ενώ τα συνολικά ποσοστά αθροίζουν στο 100% μέσα σε όλα τα κελιά του πίνακα.

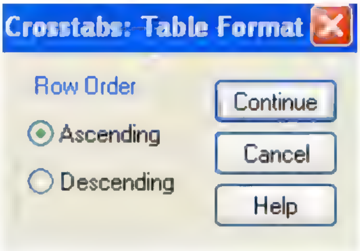
√ **Residuals:** Τα υπόλοιπα είναι οι διαφορές μεταξύ των παρατηρούμενων και αναμενόμενων τιμών του κάθε κελιού.

Στη συνέχεια, επιλέγουμε τα στατιστικά μέτρα (Statistics) (εικόνα 1.2.3) που θέλουμε να εμφανιστούν στον πίνακα συνάφειας. Μπορούμε όχι μόνο να πραγματοποιήσουμε έλεγχο ανεξαρτησίας αλλά και να αναζητήσουμε και το βαθμό και τη φύση της συνάφειας μεταξύ των μεταβλητών.



Εικόνα 1.2.3: Τα στατιστικά μέτρα της διαδικασίας Crosstabs

Τέλος, μπορούμε να καθορίσουμε τον τρόπο εμφάνισης με τη διαδικασία "Format" του πίνακα συνάφειας (εικόνα 1.2.4). Ο πίνακας μπορεί να ταξινομηθεί ως προς τη σειρά εμφάνισης (κατά αύξουσα ή φθίνουσα σειρά των διαφορετικών κατηγοριών της μεταβλητής που ορίζει τις γραμμές).



Εικόνα 1.2.4: Διαδικασία "Format"

Παράδειγμα σε ποιοτικά δεδομένα

Θα χρησιμοποιήσουμε τα δεδομένα από το αρχείο του "sales.sav". Συγκεκριμένα, το ενδιαφέρον μας αφορά την αναζήτηση σχέσης μεταξύ των μεταβλητών territory και industry, όπου είναι κωδικοποιημένες.

Επιλέγουμε διαδοχικά Analyze -> Descriptive Statistics -> Crosstabs και παίρνουμε τα αποτελέσματα που μας ενδιαφέρουν και παρουσιάζονται στους παρακάτω πίνακες.

Territory * Industry Crosstabulation

			Industry			Total
			Government	Commercial	Academic	
Territory	North	Count	137	124	121	382
		Expected Count	127,8	131,9	122,2	382,0
		% within Territory	35,9%	32,5%	31,7%	100,0%
		Adjusted Residual	1,2	-1,0	-,2	
	South	Count	112	128	105	345
		Expected Count	115,5	119,1	110,4	345,0
		% within Territory	32,5%	37,1%	30,4%	100,0%
		Adjusted Residual	-,4	1,1	-,7	
	East	Count	129	128	118	375
		Expected Count	125,5	129,5	120,0	375,0
		% within Territory	34,4%	34,1%	31,5%	100,0%
		Adjusted Residual	,4	-,2	-,3	
	West	Count	124	138	136	398
		Expected Count	133,2	137,4	127,4	398,0
		% within Territory	**31,2%**	**34,7%**	**34,2%**	100,0%
		Adjusted Residual	-1,1	,1	1,1	
Total		Count	502	518	480	1500
		Expected Count	502,0	518,0	480,0	1500,0
		% within Territory	33,5%	34,5%	32,0%	100,0%
		Adjusted Residual				

Πίνακας 1.2.1: Ο πίνακας συνάφειας των δύο μεταβλητών

Στον πίνακα 1.2.1, παρατηρούμε τα συνολικά ποσοστά της κάθε γραμμής και τα ποσοστά της κάθε υποκατηγορίας. Θα πρέπει να σημειώσουμε πως η επιλογή των ποσοστών που θα υπολογιστούν δεν είναι τυχαία. Αν κάποια από τις δύο μεταβλητές χαρακτηρίζεται σαν ανεξάρτητη θα πρέπει να υπολογίσουμε τα ποσοστά με τρόπο ώστε να αθροίζουν στο 100 για κάθε κατηγορία της ανεξάρτητης μεταβλητής.

Οι τιμές των τυποποιημένων υπολοίπων θα πρέπει να διαβάζονται σαν z-τιμές. Μεγέθη μικρότερα ή μεγαλύτερα από -2 και +2 υποδεικνύουν κελιά που διαφέρουν σαφώς από το μοντέλο της ανεξαρτησιάς.

Τα μέτρα που καταγράφουν την ένταση και τη φύση του βαθμού συνάφειας των δύο μεταβλητών χωρίζονται από το λογισμικό σε δύο κατηγορίες (Πίνακας 1.2.2 και Πίνακας 1.2.3). Στην πρώτη κατηγορία (directional–Πίνακας 1.2.2) ανήκουν εκείνα τα οποία απαιτούν το χαρακτηριστικό της μίας μεταβλητής ως ανεξάρτητης και της άλλης ως εξαρτημένης. Βασίζονται στην ιδέα ότι ο βαθμός της σχέσης που συνδέει δύο μεταβλητές μπορεί να μετρηθεί με το βαθμό στον οποίο η γνώση των τιμών της μίας μεταβλητής βελτιώνει τις προβλέψεις μας για την άλλη. Στη δεύτερη κατηγορία (Symmetric–Πίνακας 1.2.3) ανήκουν τα μέτρα για τα οποία ο διαχωρισμός αυτός δεν έχει σημασία. Οι συντελεστές Lamda και αβεβαιότητας (uncertainty coefficient) μπορούν να χρησιμοποιηθούν σαν μέτρα του βαθμού συνάφειας δύο μεταβλητών ανεξάρτητα από την κλίμακα μέτρησής τους. Στην περίπτωση όμως μεταβλητών διάταξης, η υπάρχουσα διάταξη δεν επηρεάζει σε κανένα σημείο τους υπολογισμούς για την εύρεσή τους και συνεπώς θα πρέπει να χρησιμοποιήσουμε εναλλακτικά μέτρα τα οποία θα προσδιορίζουν όχι μόνο την ένταση, αλλά και τη φύση της συνάφειας. Ο συντελεστής d του Somers είναι ένας τέτοιος συντελεστής. Αν δεν έχουμε λόγους για να θεωρήσουμε κάποια από τις δύο μεταβλητές σαν ανεξάρτητη θα πρέπει να συμβουλευτούμε τον πίνακα 1.2.3. Οι τιμές στο συγκεκριμένο παράδειγμα είναι πάρα πολύ μικρές, συνεπώς δεν έχουμε ενδείξεις για ύπαρξη συσχέτισης.

Directional Measures

			Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,013	,015	,832	,405
		Territory Dependent	,012	,015	,805	,421
		Industry Dependent	,014	,023	,615	,538
	Goodman and Kruskal tau	Territory Dependent	,001	,001		,736(c)
		Industry Dependent	,001	,001		,737(c)
	Uncertainty Coefficient	Symmetric	,001	,001	,938	,740(d)
		Territory Dependent	,001	,001	,938	,740(d)
		Industry Dependent	,001	,001	,938	,740(d)
Ordinal by Ordinal	Somers' d	Symmetric	,026	,022	1,146	,252
		Territory Dependent	,027	,024	1,146	,252
		Industry Dependent	,024	,021	1,146	,252

a.... Not assuming the null hypothesis.
b.... Using the asymptotic standard error assuming the null hypothesis.
c.... Based on chi-square approximation
d.... Likelihood ratio chi-square probability.

Πίνακας 1.2.2: Μέτρα της έντασης και της φύσης του βαθμού συνάφειας των μεταβλητών

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	,026	,022	1,146	,252
	Kendall's tau-c	,027	,024	1,146	,252
	Gamma	,036	,032	1,146	,252
	Spearman Correlation	,030	,026	1,153	,249(c)
Interval by Interval	Pearson's R	,029	,026	1,140	,254(c)
N of Valid Cases		1500			

a.... Not assuming the null hypothesis.
b.... Using the asymptotic standard error assuming the null hypothesis.
c.... Based on normal approximation.

Πίνακας 1.2.3: Μέτρα της έντασης και της φύσης του βαθμού συνάφειας των μεταβλητών.

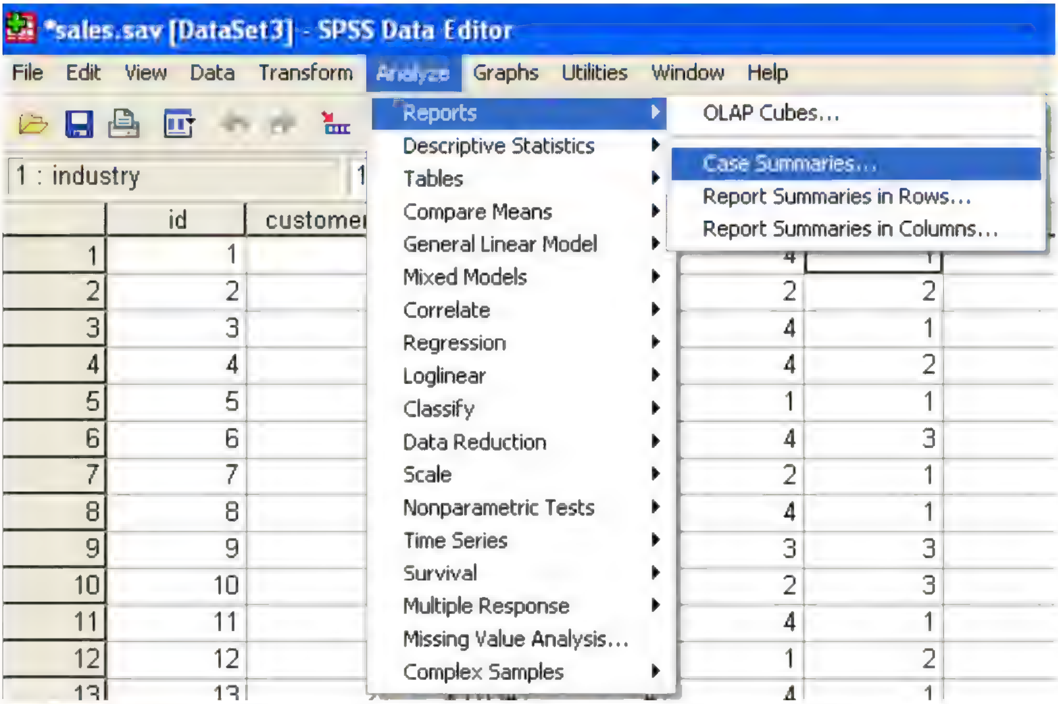
Οι τεχνικές που μπορούμε να υιοθετήσουμε για την Περιγραφική Στατιστική στην περίπτωση των ποσοτικών μεταβλητών, είναι ιδιαίτερα πλούσιες. Γενικά μπορούμε να πραγματοποιήσουμε:

••• Παρουσίαση (αλγεβρική ή και γραφική) των στατιστικών μέτρων μιας ποσοτικής μεταβλητής, μέσα στις διάφορες κατηγορίες κάποιας ποιοτικής.

••• Παρουσιάση της μεταβολής μιας ποσοτικής μεταβλητής σε σχέση με τη μεταβολή κάποιας άλλης ποσοτικής και αναζήτηση (πιθανής) μεταξύ τους σχέσης.

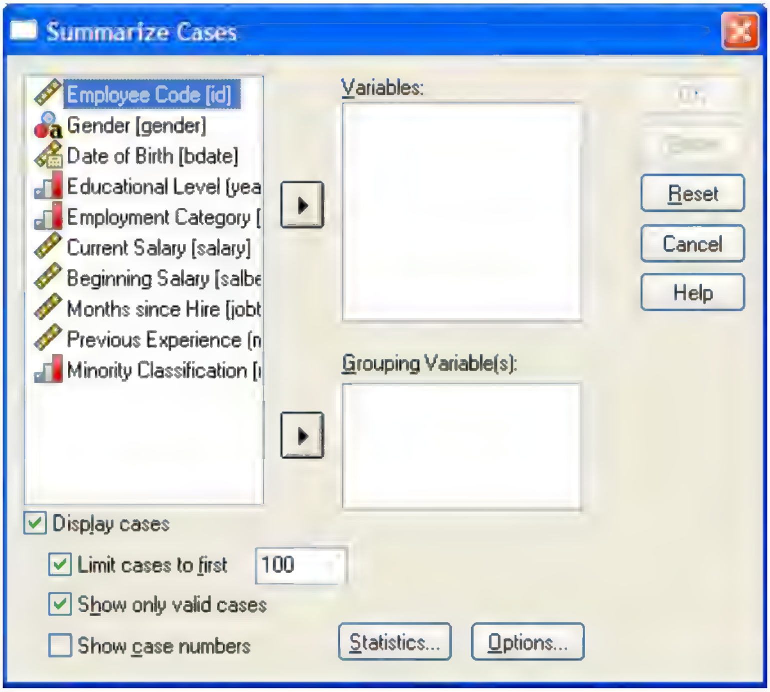
Σε αυτό το σημείο το ενδιαφέρον θα περιοριστεί απλά στην εύρεση μερικών αριθμητικών περιγραφικών μέτρων μιας ποσοτικής μεταβλητής μέσα στις διάφορες κατηγορίες κάποιας ποιοτικής. Μπορούμε να χρησιμοποιήσουμε τη διαδικασία "Case Summaries" η οποία θα μας δώσει επιπλέον μια εικόνα του συνόλου των παρατηρήσεων.

Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγουμε:
Analyze => Reports => Case Summaries



Εικόνα 1.2.5: Διαδικασία Case Summaries από τη βασική ράβδο.

Το πλαίσιο διαλόγου στο SPSS για την πραγματοποίηση Περιγραφικής Στατιστικής, στις τιμές μιας ποσοτικής μεταβλητής μέσα στις διάφορες κατηγορίες κάποιας ποιοτικής μεταβλητής με τη βοήθεια της διαδικασίας "Case Summaries", παρουσιάζεται παρακάτω.



Εικόνα 1.2.6: Διαδικασία "Case Summaries"

Στα επόμενα κεφάλαια αναλύονται τα παραπάνω λεπτομερώς, χρησιμοποιώντας την κατάλληλη στατιστική τεχνική, για να έχουμε μια ισχυρή στατιστική συμπερασματολογια.

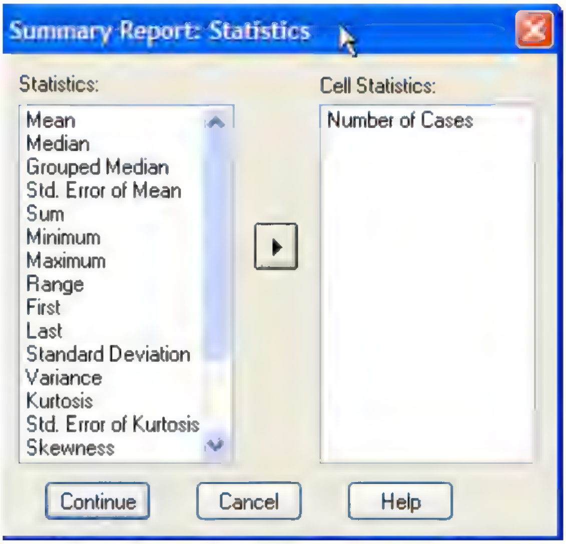
--• Διαλέγουμε την ποσοτική μεταβλητή (ή τις μεταβλητές) που θέλουμε να μελετήσουμε και τη μετακινούμε στο παράθυρο «Variables».

--• Ορίζουμε τη μεταβλητή που θα διαχωρίσει τις τιμές του δείγματος που έχουμε και τη μετακινούμε στο παράθυρο Grouping Variable(s). Η μεταβλητή αυτή πρέπει να είναι ποιοτική. Μπορούμε να επιλέξουμε ταυτόχρονα περισσότερες από μία μεταβλητές.

Στην περίπτωση αυτή, πραγματοποιούνται οι υπολογισμοί που θα ζητηθούν για τον κάθε δυνατό συνδυασμό κατηγοριών, των εμπλεκόμενων μεταβλητών που εμείς έχουμε ορίσει.

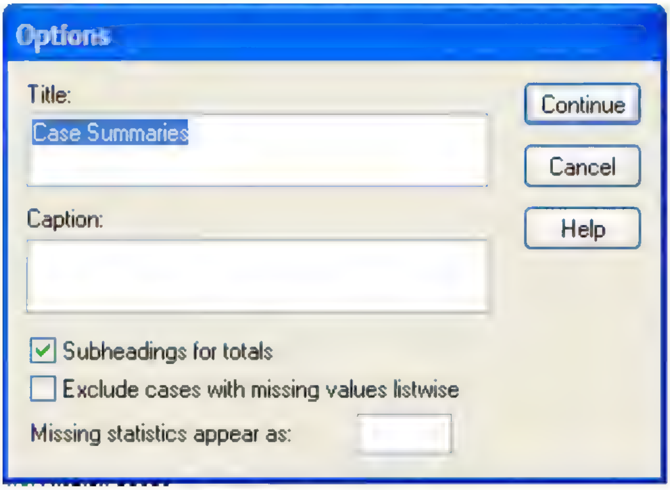
--• Καθορίζουμε το πλήθος των περιπτώσεων, που θα απαριθμούν μέσα σε κάθε δείγμα (Display Cases). Το λογισμικό θα εμφανίσει τις εκατό πρώτες παρατηρήσεις του αρχείου των δεδομένων.

--• Αν επιθυμούμε επιπλέον και κάποια από τα αριθμητικά περιγραφικά μέτρα, θα πρέπει να ενεργοποιήσουμε την επιλογή "Statistics" (Εικόνα 1.2.7).



Εικόνα 1.2.7: Διαδικασία "Statistics"

Από τη λειτουργία "Options" στην Εικόνα 1.2.8 καθορίζουμε τον τρόπο εμφάνισης των αποτελεσμάτων.



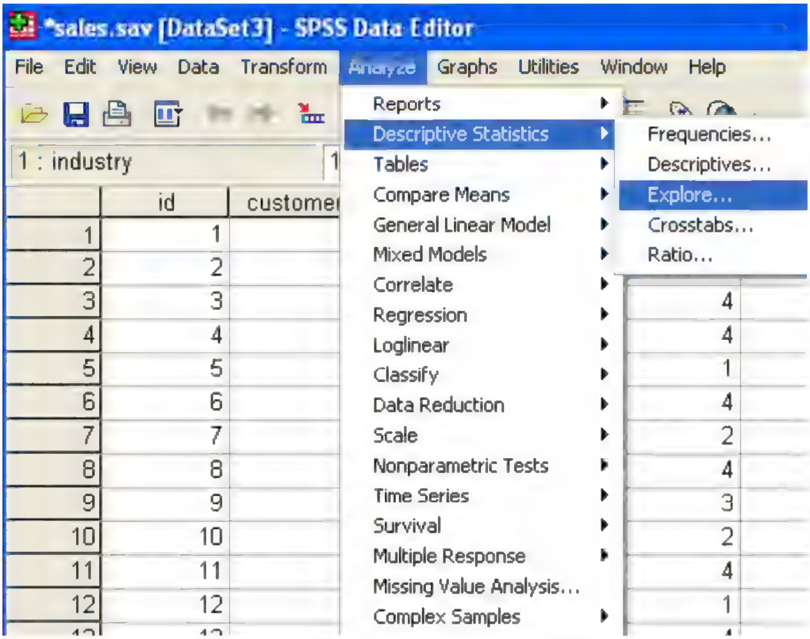
Εικόνα 1.2.8: Διαδικασία "Options"

Στο συγκεκριμένο πλαίσιο μπορούμε να ορίσουμε τον τίτλο (Title) των αποτελεσμάτων, τη λεζάντα (Caption) των αποτελεσμάτων και αν θέλουμε να εμφανιστούν ή όχι οι υπότιτλοι (Subheading for totals). Επίσης, μπορούμε να χρησιμοποιήσουμε στους υπολογισμούς μόνο τις περιπτώσεις που είναι ταυτόχρονα έγκυρες για όλες τις μεταβλητές των καταλόγων (Exclude cases with missing values listwise) και να δηλώσουμε με κάποιο χαρακτηριστικό τρόπο το στατιστικό μέτρο που δεν μπορεί να υπολογιστεί με έναν αστερίσκο, κάποια λέξη κ.λπ. (missing statistics appear as).

Επιπλέον με τη διαδικασία "Explore" μπορούμε να πετύχουμε την πιο πλούσια και πλήρη περιγραφική στατιστική των παρατηρήσεων μιας ποσοτικής μεταβλητής μέσα στις διάφορες κατηγορίες κάποιας ποιοτικής.

Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγουμε:

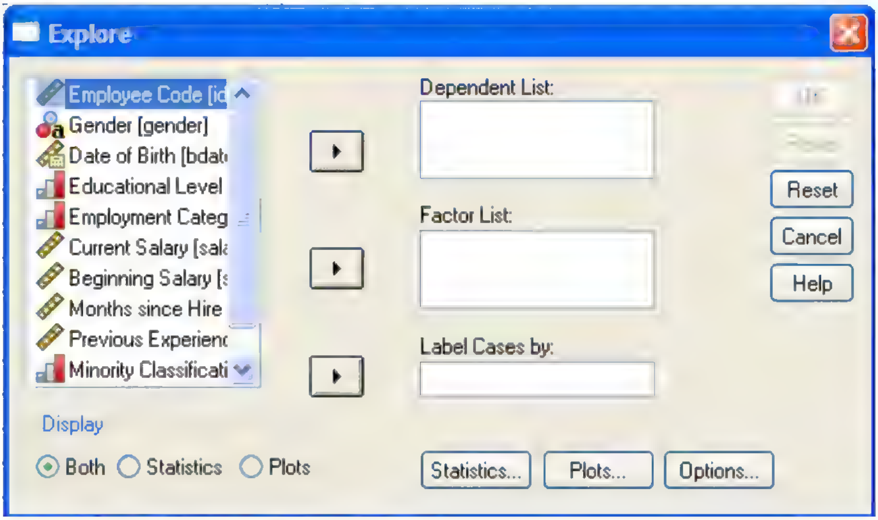
Analyze => Descriptives Statistics => Explore



Εικόνα 1.2.9: Διαδικασία Explore από τη βασική ράβδο.

Στα επόμενα κεφάλαια αναλύονται τα παραπάνω λεπτομερώς, χρησιμοποιώντας την κατάλληλη στατιστική τεχνική, για να έχουμε μια ισχυρή στατιστική συμπερασματολογια.

Το πλαίσιο διαλόγου, που εμφανίζεται στο SPSS για την πραγματοποίηση Περιγραφικής Στατιστικής των τιμών μιας ποσοτικής μεταβλητής μέσα στις διάφορες κατηγορίες κάποιας ποιοτικής χρησιμοποιώντας τη διαδικασία «Explore» είναι το ακόλουθο.



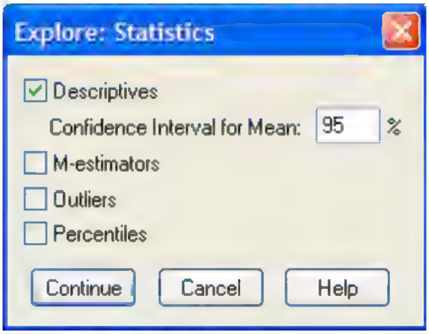
Εικόνα 1.2.10: Διαδικασία "Explore"

• Διαλέγουμε τη μεταβλητή (ή τις μεταβλητές), που θέλουμε να περιγράψουμε και τη μετακινούμε στο παράθυρο Dependent List.

• Ορίζουμε τη μεταβλητή, που θα διαχωρίσει τις τιμές του δείγματος που έχουμε και τη μετακινούμε στο παράθυρο Factor List.

• Στο Label Cases by, τοποθετούμε τη μεταβλητή ταυτοποίησης των παράτυπων σημείων, που εξ ορισμού είναι ο αύξων αριθμός της κάθε περίπτωσης.

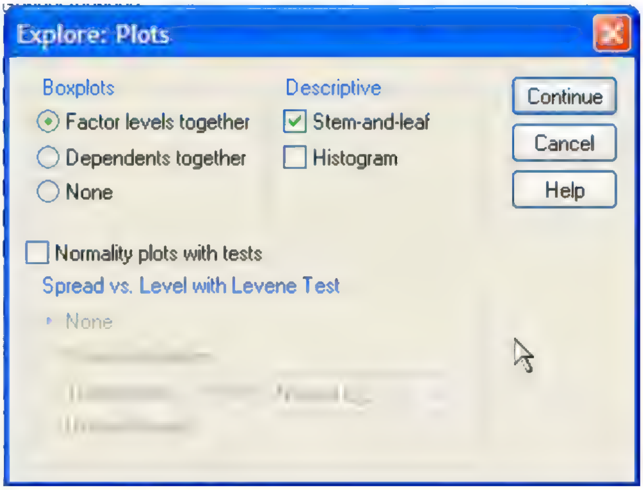
• Αν επιθυμούμε να υπολογίσουμε περισσότερα από τα γνωστά στατιστικά μέτρα, ενεργοποιούμε την καρτέλα "Statistics" (Εικόνα 1.2.11).



Εικόνα 1.2.11: Διαδικασία "Statistics"

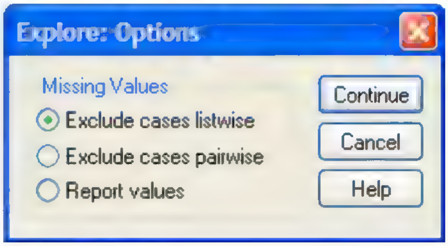
Στην κατηγορία "Descriptives" περιλαμβάνονται τα βασικά στατιστικά μέτρα. Στην περίπτωση, που έχουμε πολλά ακραία σημεία, το λογισμικό διαθέτει τέσσερις ανθεκτικούς εκτιμητές μεγίστης πιθανοφάνειας της κεντρικής τάσης (M-estimators). Με την επιλογή "Outliers" το SPSS θα υποδείξει τις πέντε μεγαλύτερες και τις πέντε μικρότερες τιμές (δεν είναι σίγουρα ακραία ή παράτυπα σημεία). Τέλος, με την επιλογή "Percentiles", πραγματοποιείται ο υπολογισμός του 5ου, 10ου, 25ου, 50ου, 75ου, 90ου και 95ου εκατοστιαίου σημείου.

Με την επιλογή "Plots" (Εικόνα 1.2.12) μπορούμε να ενεργοποιήσουμε τις γραφικές παραστάσεις που επιθυμούμε.



Εικόνα 1.2.12: Διαδικασία "Plots"

Επίσης, από την επιλογή "Options" (Εικόνα 1.2.13) μπορούμε να χειριστούμε τις ακραίες ή παράτυπες τιμές στο σύνολο των δεδομένων, που επεξεργαζόμαστε.



Εικόνα 1.2.13: Διαδικασία "Options"

1.3 Ανάλυση της εισαγωγής δεδομένων στο SPSS από Aschii, excel, dbase, access αρχεία δεδομένων.

Το SPSS είναι σε θέση να χειριστεί αρχεία δεδομένων που έχουν δημιουργηθεί από τα πιο διαδεδομένα λογισμικά, όπως το Excel (έκδοση 4 ή προγενέστερη), Lotus 1-2-3, dBASE κ.λπ. Μπορεί ακόμη να αναγνωρίσει διάφορες μορφές ASCHII αρχείων και φυσικά αρχεία που δημιουργήθηκαν από το ίδιο. Τα αρχεία αυτά, αναγνωρίζονται εύκολα από το επίθεμα ".sav" που το λογισμικό αυτόματα προσθέτει στο όνομα τους.

Ανάγνωση αρχείων δεδομένων (SPSS)
Η ανάγνωση ενός αρχείου δεδομένων, όταν αυτά δημιουργήθηκαν από το ίδιο το λογισμικό, είναι απλή υπόθεση για το SPSS. Από τη βασική ράβδο προτιμήσεων (menus) διαδοχικά επιλέγουμε Files, Open. Εξ' ορισμού, το λογισμικό εμφανίζει όλα τα SPSS αρχεία δεδομένων (*.sav) που βρίσκονται στον κατάλογο (directory) που υποδείχθηκε. Μετά την επιλογή του αρχείου που μας ενδιαφέρει, με το πλήκτρο ΟΚ μεταφέρουμε τα δεδομένα του στον SPSS Data Editor οπότε μπορούμε να προχωρήσουμε στην όποια στατιστική ανάλυση επιθυμούμε.

Ανάγνωση αρχείων δεδομένων (ASCHII)
Αν τα δεδομένα μας είναι αποθηκευμένα σε ένα απλό αρχείο κειμένου μπορούμε να τα μεταφέρουμε στο SPSS και να ορίσουμε ονόματα για τις μεταβλητές που περιέχουν. Εν γένει, τα δεδομένα μπορεί να έχουν καταγραφεί με δύο μορφές: τη συστηματική (fixed), όπου η τιμή της κάθε μεταβλητής βρίσκεται στην ίδια πάντοτε στήλη ή στήλες, και την ελεύθερη (free-field), όπου οι τιμές των διαφόρων μεταβλητών ξεχωρίζουν μόνον από την ύπαρξη ενός κενού ή κόμματος μεταξύ τους.

Η συστηματική μορφή. Κύριο γνώρισμα εδώ είναι ότι η τιμή της κάθε μεταβλητής βρίσκεται στην ίδια πάντοτε θέση, δηλαδή στην ίδια στήλη σε όλες τις γραμμές (records).

Από τη βασική ράβδο προτιμήσεων (menus) επιλέγουμε Files και συναντάμε τα ακόλουθα:

- **Browse:** Επιλέγουμε το αρχείο (.dat) που περιέχει τα στοιχεία που μας ενδιαφέρουν.

- **Defined Variables:** Για κάθε μεταβλητή προσδιορίζουμε ένα όνομα (Name), τη στήλη από την οποία ξεκινά η τιμή της (Start Column), εκείνη που τελειώνει (End Column) και τέλος τον τύπο της (Data Type). Η ενεργοποίηση της όποιας επιλογής μας γίνεται με το πλήκτρο Add.

- **Γενικοί κανόνες:** Δεν έχει σημασία η σειρά με την οποία θα δηλωθούν οι μεταβλητές. Το λογισμικό θα τις ταξινομήσει κατά σειρά και στήλη από την οποία ξεκινούν, χωρίς κανένα πρόβλημα.

Δεν είναι απαραίτητο να διαβάσουμε ή να μεταφέρουμε στο SPSS όλα τα στοιχεία/ μεταβλητές που βρίσκονται στο (ASCHII) αρχείο δεδομένων. Η όλη διαδικασία θα διαβάσει και θα μεταφέρει στο λογισμικό μόνον τα στοιχεία που βρίσκονται στις στήλες ή/και σειρές που ορίστηκαν, παραλείποντας όλες τις υπόλοιπες. Τέλος μπορούμε για τις ίδιες θέσεις-στήλες να αντιστοιχίσουμε περισσότερες από μία SPSS μεταβλητές.

Κανόνες ονομασίας των μεταβλητών. Ως όνομα μεταβλητής μπορεί να χρησιμοποιηθεί οποιοσδήποτε συνδυασμός γραμμάτων του αγγλικού αλφαβήτου, αριθμών και των ειδικών συμβόλων: @, ., #, _ , $. Ο πρώτος χαρακτήρας στην επιχειρούμενη ονοματολογία πρέπει να είναι γράμμα, ο τελευταίος δεν μπορεί να είναι η τελεία (.), ενώ το όλο μέγεθος του ονόματος δεν πρέπει να ξεπερνά τους οκτώ χαρακτήρες. Φυσικά δεν μπορούμε να έχουμε δύο διαφορετικές μεταβλητές με το ίδιο όνομα, ενώ δεν υπάρχει διάκριση μεταξύ κεφαλαίων και πεζών γραμμάτων.

Η ελεύθερη μορφή. Εδώ οι μεταβλητές έχουν καταγραφεί με την ίδια σειρά για κάθε περίπτωση (case), αλλά οι τιμές τους δεν βρίσκονται στην ίδια θέση. Οι τιμές των διαφόρων μεταβλητών ξεχωρίζουν από την ύπαρξη ενός κενού ή κόμματος μεταξύ τους, ενώ σε μια γραμμή (record) μπορεί να υπάρχουν περισσότερες από μία περιπτώσεις.

Είναι σημαντικό οι μεταβλητές να δηλωθούν με τη σειρά που καταγράφηκαν στο (ASCHII) αρχείο δεδομένων. Κάθε νέος ορισμός μεταβλητής προστίθεται στη συνέχεια από τους ήδη υπάρχοντες και το SPSS θα διαβάσει τα στοιχεία με αυτή τη σειρά. Επιπλέον, θα πρέπει να ορίσουμε όλες τις υπάρχουσες μεταβλητές. Το λογισμικό καθορίζει το τέλος της μιας περίπτωσης (case) και την αρχή της άλλης αποκλειστικά και μόνο από τον αριθμό των μεταβλητών που δηλώθηκαν.

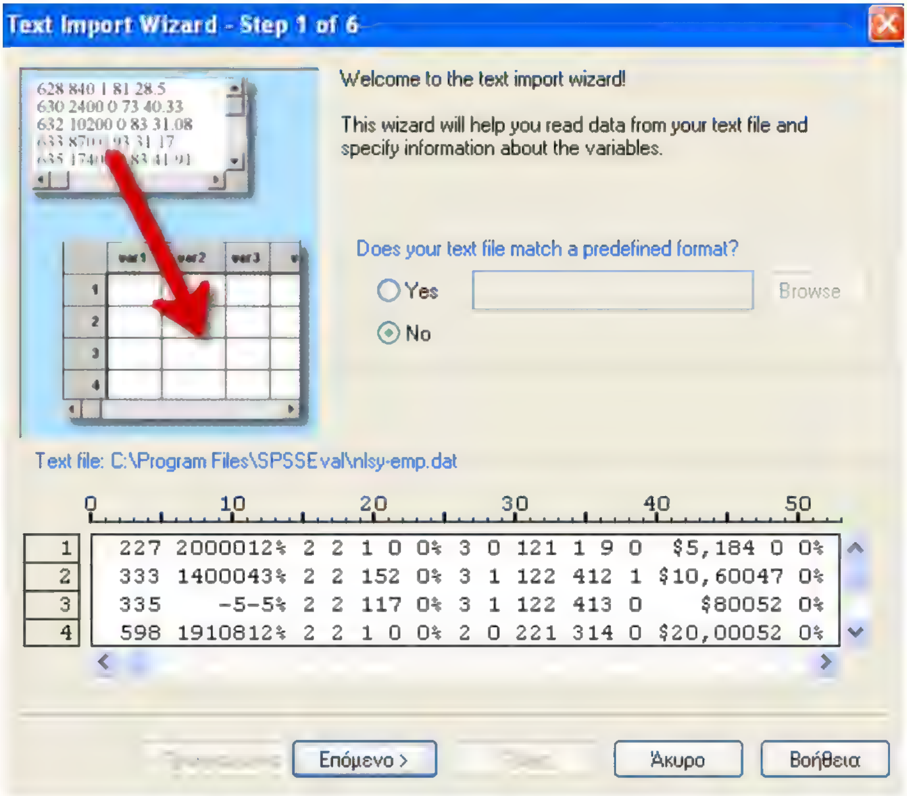
Ανάγνωση αρχείων δεδομένων (διάφορα formats)

Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα αρχεία που δημιουργήθηκαν από ένα λογιστικό φύλλο (π.χ. Excel) και την dBase.

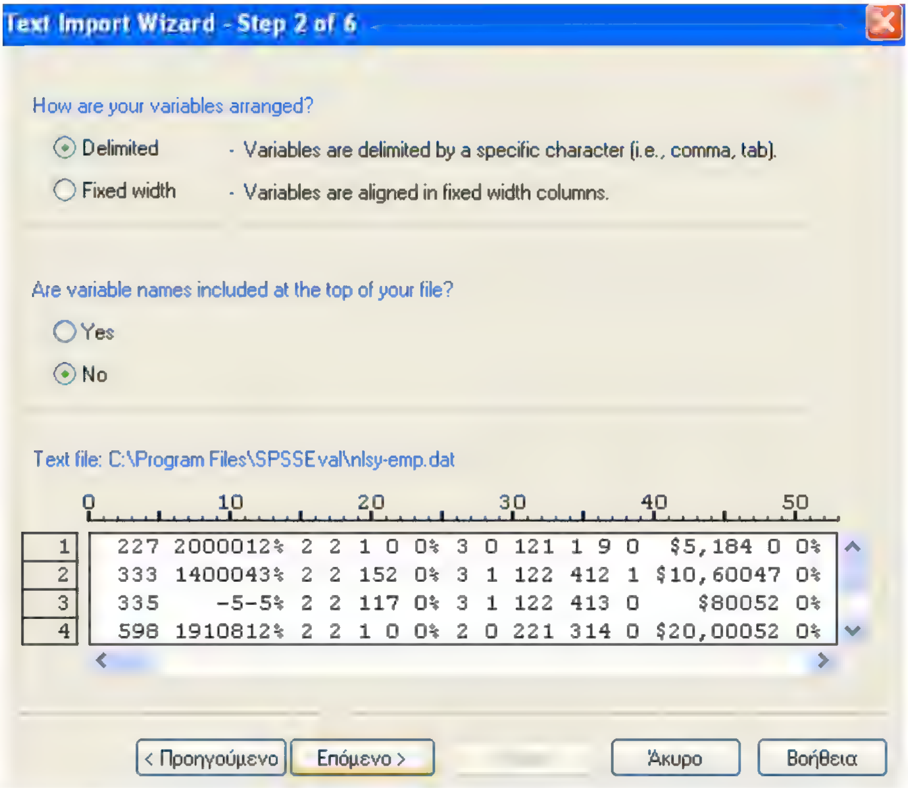
Λογιστικά φύλλα. Οι κανόνες που ισχύουν είναι μάλλον απλοί. Οι στήλες χαρακτηρίζονται σαν μεταβλητές και οι γραμμές σαν περιπτώσεις. Αν στο λογιστικό φύλλο περιέχονται τα ονόματα των μεταβλητών, αυτά θα μεταφερθούν και στο SPSS. Διαφορετικά οι μεταβλητές θα ονομαστούν σειριακά. Στην περίπτωση που υπάρχουν κενά κελιά, αν αυτά αντιστοιχούν σε αριθμητικές μεταβλητές θα γίνουν ελλείπουσες τιμές (missing values), ενώ αν αντιστοιχούν σε αλφαριθμητικές θα παραμείνουν κενά κελιά και στο SPSS.

dBASE. Ούτε εδώ υπάρχουν δυσκολίες ή προβλήματα. Τα ονόματα των πεδίων (fields) γίνονται ονόματα μεταβλητών στο SPSS.

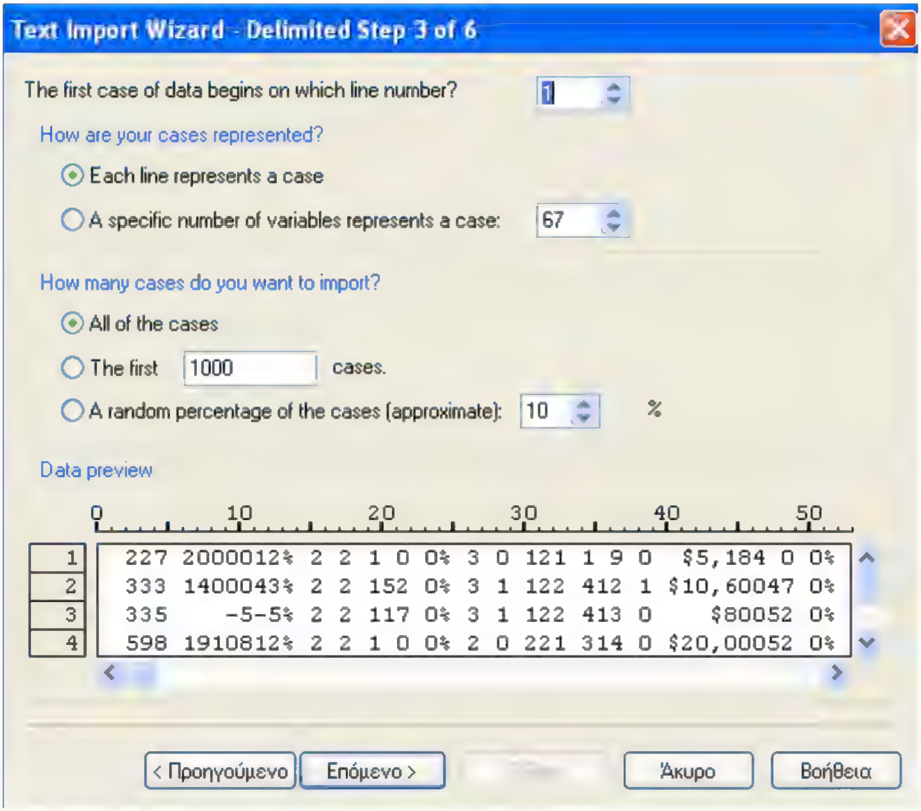
Παρακάτω, επισημαίνουμε κατά σειρά τα output, που εμφανίζονται στο SPSS, για την εισαγωγή κάθε φορά των δεδομένων από τα παραπάνω είδη αρχείων δεδομένων. Σε κάθε βήμα μπορούμε να επέμβουμε ανάλογα με τη μορφή των αρχικών μας δεδομένων και τον τρόπο που θα ήταν σκόπιμο να παρουσιαστούν στο SPSS.



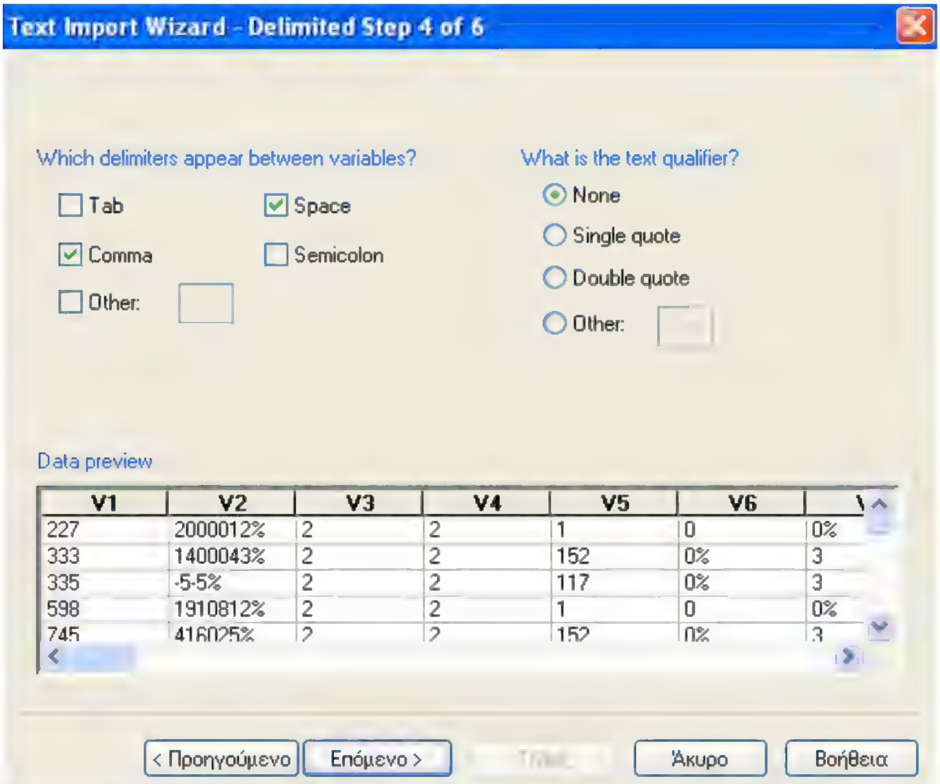
Εικόνα 1.3.1: 1ο Βήμα για ανάγνωση αρχείων δεδομένων



Εικόνα 1.3.2: 2ο Βήμα για ανάγνωση αρχείων δεδομένων

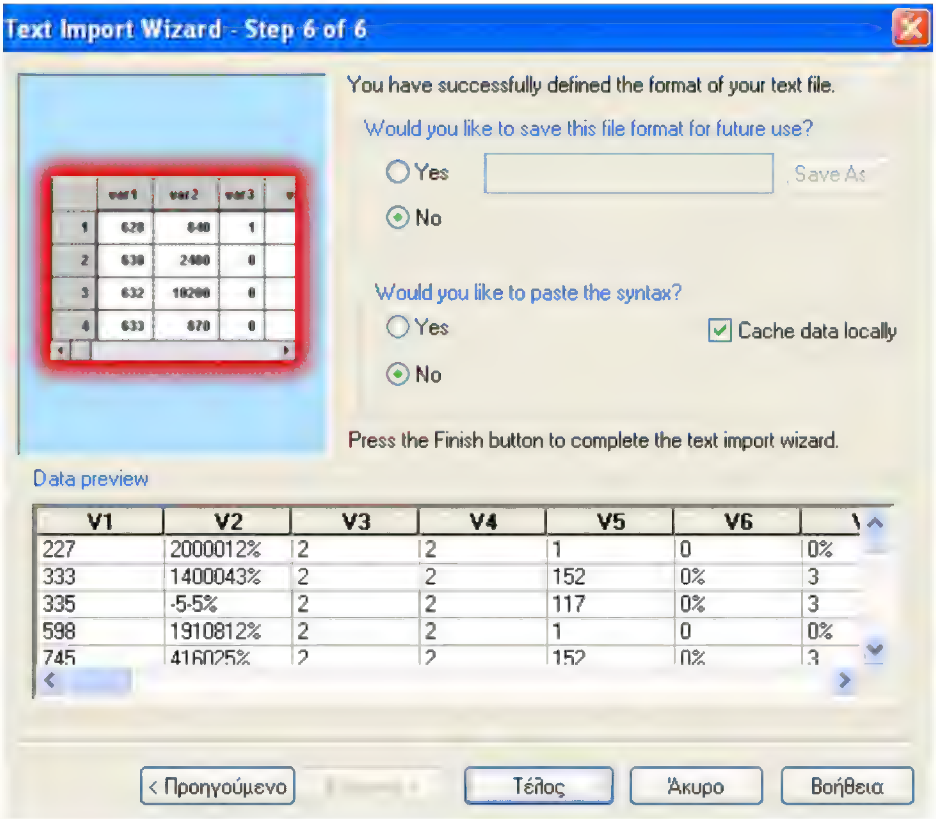


Εικόνα 1.3.3: 3ο Βήμα για ανάγνωση αρχείων δεδομένων



Εικόνα 1.3.4: 4ο Βήμα για ανάγνωση αρχείων δεδομένων

Εικόνα 1.3.5: 5ο Βήμα για ανάγνωση αρχείων δεδομένων



Εικόνα 1.3.6: 6ο Βήμα για ανάγνωση αρχείων δεδομένων

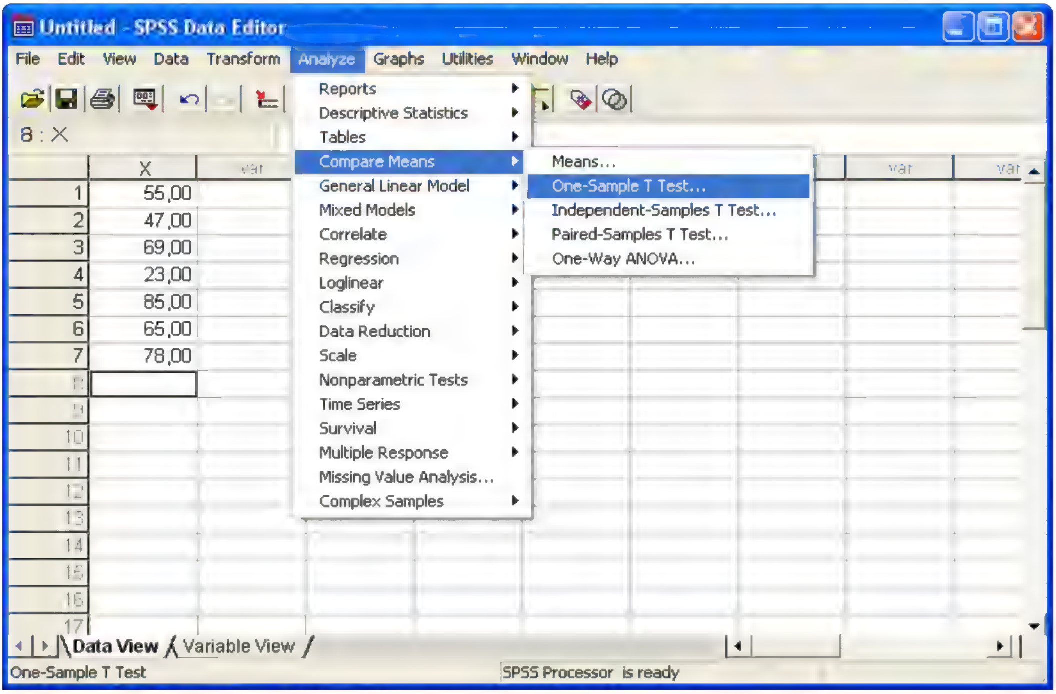
Αποθήκευση αρχείων δεδομένων

Για να αποθηκεύσουμε αλλαγές που πραγματοποιήσαμε σε ένα υπάρχον αρχείο δεδομένων του SPSS από τη βασική ράβδο προτιμήσεων διαδοχικά επιλέγουμε **File -> Save** τότε τα νέα δεδομένα αντικαθιστούν τα παλαιότερα.

Στην περίπτωση που θέλουμε να αποθηκεύσουμε ένα καινούργιο αρχείο δεδομένων θα πρέπει να επιλέξουμε **File -> Save as** και στη συνέχεια να δώσουμε το όνομα που επιθυμούμε.

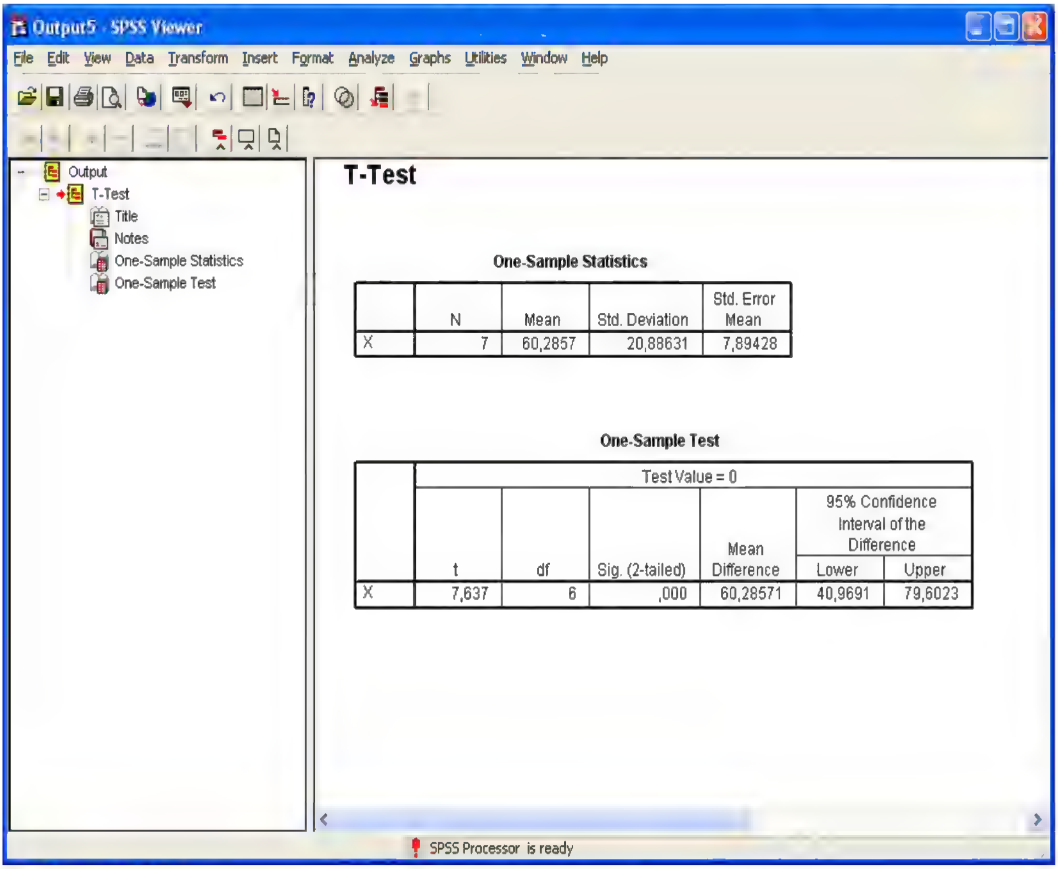
1.4 Διάστημα εμπιστοσύνης με χρήση SPSS

Μπορούμε πολύ απλά να κατασκευάσουμε ένα διάστημα εμπιστοσύνης χρησιμοποιώντας το SPSS. Αρχικά καταχωρούμε τα δεδομένα μας σε μια μεταβλητή, έστω X. Επιλέγουμε:



Εικόνα 1.4.1: Επιλογή διαστήματος εμπιστοσύνης

Επιλέγουμε τη μεταβλητή για την οποία θέλουμε να κατασκευάσουμε το διάστημα εμπιστοσύνης και στην καρτέλα Options καθορίζουμε το επίπεδο σημαντικότητας του διαστήματος αυτού (από το σύστημα έχει οριστεί ως 95%). Το αποτέλεσμα που παίρνουμε είναι:



T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
X	7	60,2857	20,88631	7,89428

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
X	7,637	6	,000	60,28571	40,9691	79,6023

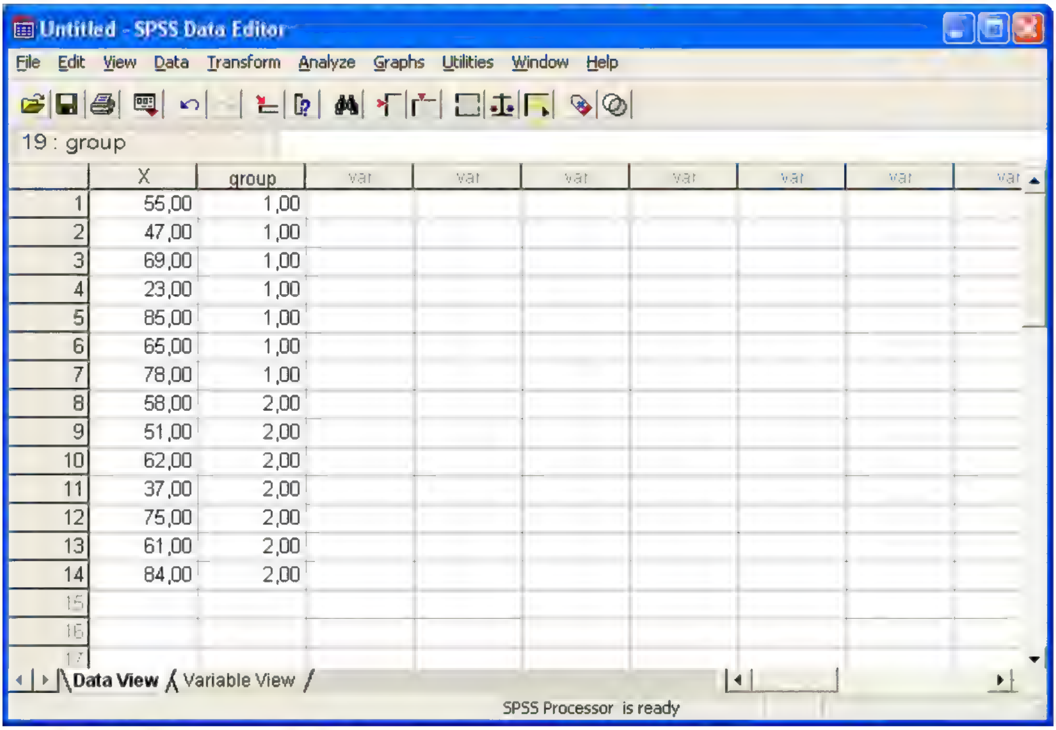
Εικόνα 1.4.2: Κατασκευή διαστήματος εμπιστοσύνης

Ένα 95% διάστημα εμπιστοσύνης έχει οριστεί το (40.9691 , 79.6023). Προφανώς όσο πιο ισχυρό θέλουμε να είναι το διάστημα εμπιστοσύνης τόσο πιο ευρύ θα είναι. Δηλαδή, ένα 99% διάστημα εμπιστοσύνης θα είχε όρια (31.02 , 89.55).

Για τα δεδομένα του παραδείγματος να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης.

X: 55 47 69 23 85 65 78
Y: 58 51 62 37 75 61 84

Υποθέτουμε ότι έχουμε ανεξάρτητους πληθυσμούς. Καταχωρούμε τα δεδομένα ως εξής:

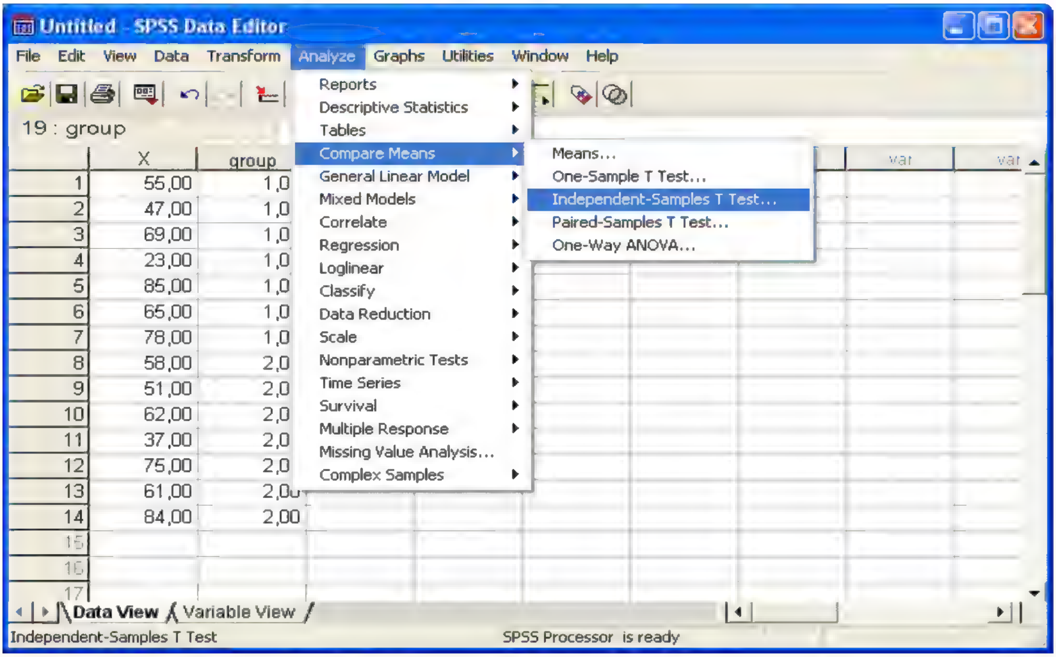


	X	group
1	55,00	1,00
2	47,00	1,00
3	69,00	1,00
4	23,00	1,00
5	85,00	1,00
6	65,00	1,00
7	78,00	1,00
8	58,00	2,00
9	51,00	2,00
10	62,00	2,00
11	37,00	2,00
12	75,00	2,00
13	61,00	2,00
14	84,00	2,00

Εικόνα 1.4.3: Καταχώριση δεδομένων

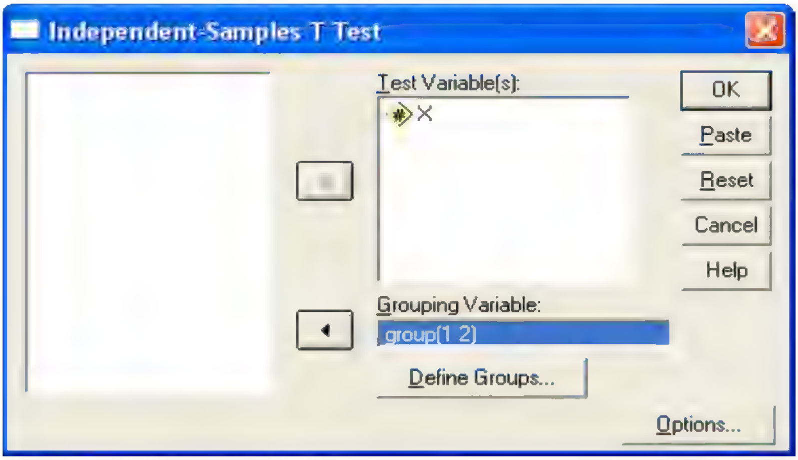
Στη μεταβλητή **group** το 1 αντιστοιχεί στα δεδομένα του X δείγματος και το 2 στα δεδομένα του Y δείγματος.

Στη συνέχεια επιλέγουμε:



Εικόνα 1.4.4: Επιλογή Δ.Ε. για διαφορά μέσων τιμών

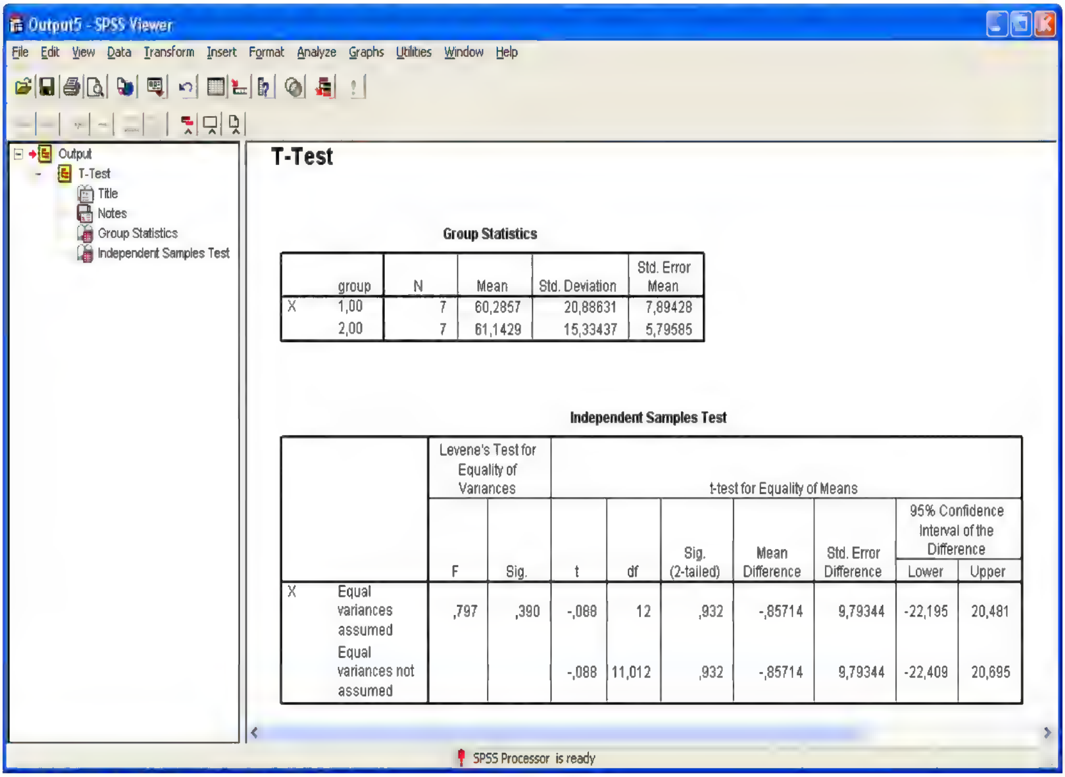
Στο κεντρικό menu ακολουθούμε την εξής διαδικασία:



Εικόνα 1.4.5: Κεντρικό menu Δ.Ε.

Στο πεδίο **Grouping Variable** ορίσαμε την κατηγοριοποίηση της μεταβλητής

Τα αποτελέσματα φαίνονται παρακάτω:



Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
X	1,00	7	60,2857	20,88631	7,89428
	2,00	7	61,1429	15,33437	5,79585

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
X	Equal variances assumed	,797	,390	-,088	12	,932	-,85714	9,79344	-22,195	20,481
	Equal variances not assumed			-,088	11,012	,932	-,85714	9,79344	-22,409	20,695

Εικόνα 1.4.6: Αποτέλεσμα Δ.Ε.

Το SPSS μας δίνει και για τις δύο υποθέσεις τα αποτελέσματα σε ένα κοινό test. Μάλιστα μέσω του test του Levene μας καθοδηγεί για το αν πρέπει να υποθέσουμε ότι έχουμε ίσες ή άνισες διακυμάνσεις.

Στο παράδειγμα μας δεχόμαστε την υπόθεση περί ίσων διακυμάνσεων αφού το p-value έχει τιμή > 0.05. Έτσι, ένα 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών είναι το (-22.195 , 20.481).

Αν υποθέταμε ότι οι διακυμάνσεις ήταν άνισες το αντίστοιχο διάστημα εμπιστοσύνης θα ήταν το (-22.409 , 20.695).

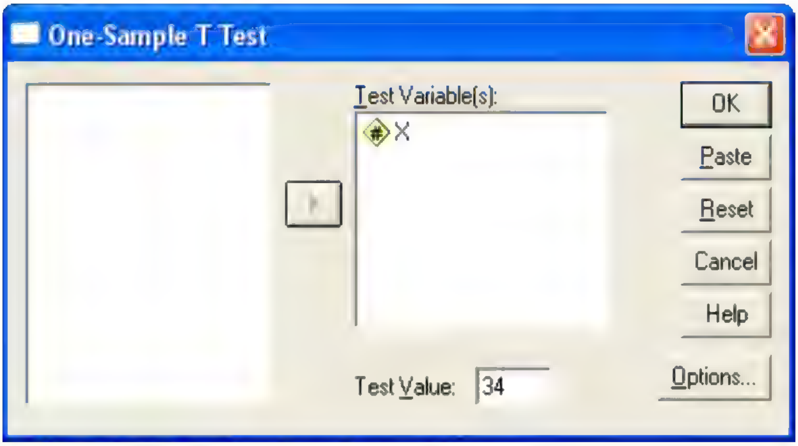
1.5 Έλεγχος Υποθέσεων με χρήση SPSS

Για τα δεδομένα του παραδείγματος να ελεγχθεί η υπόθεση ότι η μέση τιμή είναι μικρότερη από 34:

21 51 32 29 42 47 36

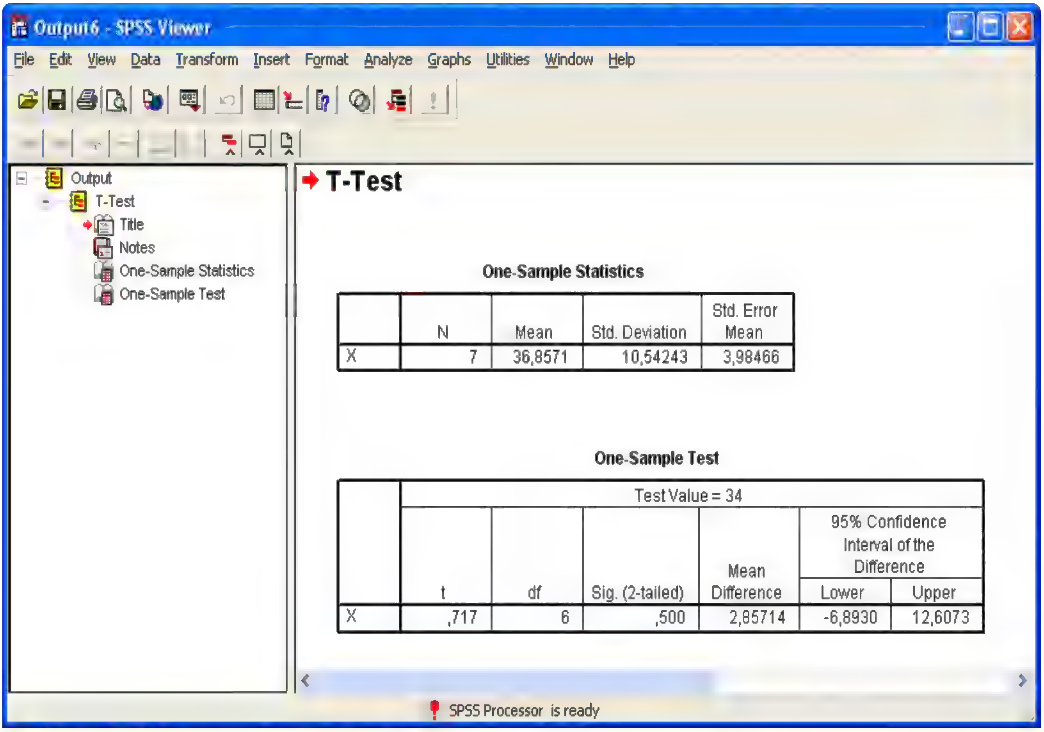
Θέλουμε δηλαδή να πραγματοποιήσουμε τον έλεγχο της μορφής:
$H_0: \mu = 34$
$H_1: \mu < 34$

Αφού καταχωρήσουμε τα δεδομένα ακριβώς όπως στην περίπτωση των διαστημάτων εμπιστοσύνης στο κεντρικό πλαίσιο (καρτέλα test value) ορίζουμε την τιμή που θέλουμε να γίνει ο έλεγχος.



Εικόνα 1.5.1: Επιλογή τιμής για έλεγχο

Το αποτέλεσμα που παίρνουμε φαίνεται στην παρακάτω εικόνα.



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
X	7	36,8571	10,54243	3,98466

One-Sample Test

	Test Value = 34					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
X	,717	6	,500	2,85714	-6,8930	12,6073

Εικόνα 1.5.2: Αποτέλεσμα ελέγχου

Αρχικά βλέπουμε ότι η μέση τιμή για το δείγμα μας ήταν 36.8, τιμή όχι πολύ μακριά από αυτήν που θέλαμε να ελέγξουμε. Το p-value δίνει τιμή 0.5 για τον αμφίπλευρο έλεγχο. Εμείς έχουμε μονόπλευρο και επομένως το p-value έχει τιμή μεγαλύτερη από 0.05, συνεπώς δεν απορρίπτουμε τη μηδενική υπόθεση. Άρα δε μπορούμε να θεωρήσουμε ότι η μέση τιμή είναι μικρότερη από 34.

ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΜΕ ΧΡΗΣΗ SPSS

Για τα δεδομένα του παραδείγματος να ελεγχθεί η υπόθεση ότι οι μέσες τιμές των δειγμάτων διαφέρουν μεταξύ τους.

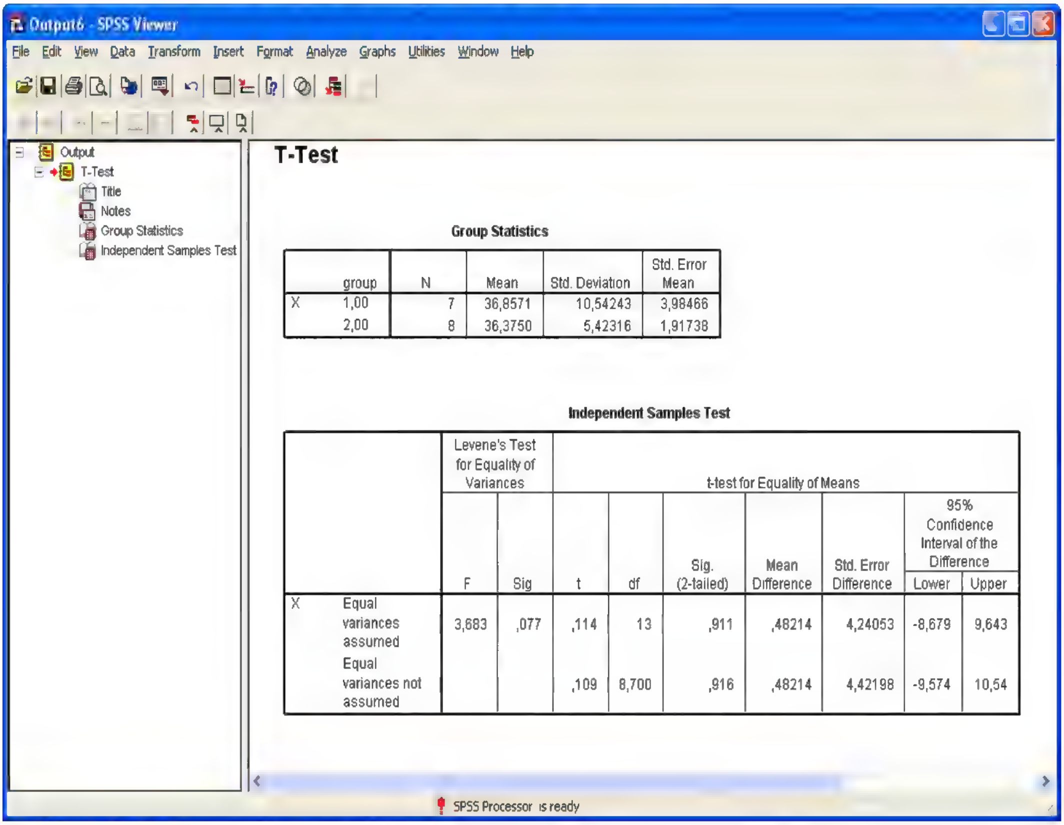
X: 21 51 32 29 42 47 36
Y: 28 44 40 35 37 30 36 41

Θέλουμε να πραγματοποιήσουμε έναν αμφίπλευρο έλεγχο της μορφής:

$H_0: \mu = \mu_X - \mu_Y = 0$
$H_1: \mu = \mu_X - \mu_Y \neq 0$

Καταχωρούμε τα δεδομένα ακριβώς όπως και στην περίπτωση του διαστήματος εμπιστοσύνης καθορίζοντας παράλληλα την κατηγορία ανάλογα με την προέλευση τους δείγματος (1 για το X και 2 για το Y).

Τα αποτελέσματα που δίνει το SPSS είναι τα ακόλουθα:



Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
X	1,00	7	36,8571	10,54243	3,98466
	2,00	8	36,3750	5,42316	1,91738

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
X	Equal variances assumed	3,683	,077	,114	13	,911	,48214	4,24053	-8,679	9,643
	Equal variances not assumed			,109	8,700	,916	,48214	4,42198	-9,574	10,54

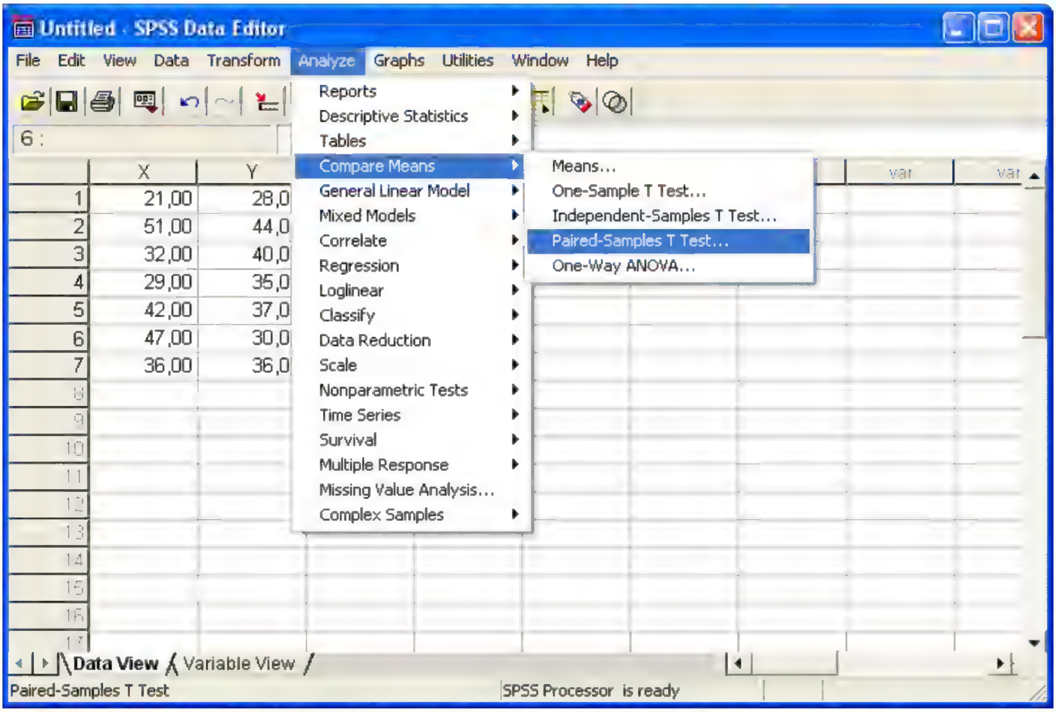
Εικόνα 1.5.3: Αποτέλεσμα ελέγχου

Αρχικά με το test του Levene θα καθοριστεί αν μπορούμε να θεωρήσουμε ίσες τις διακυμάνσεις ή όχι. Το p-value έχει τιμή 0.077, τιμή μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση περί ισότητας των διακυμάνσεων. Άρα ανήκουμε στην περίπτωση ίσων διακυμάνσεων.

Για τον έλεγχο μας τώρα περί ισότητας των δύο μέσων, το p-value δίνει τιμή ίση με 0.911 η οποία και είναι σαφώς μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε τη μηδενική ισότητα περί ισότητας των δύο μέσων τιμών. Να σημειώσουμε ότι κάτι τέτοιο φαινόταν αναμενόμενο από τη στιγμή που το ένα δείγμα είχε μέση τιμή ίση με 36.9 και το άλλο 36.4.

Τέλος, βλέπουμε και ένα 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών που είναι (-8.679 , 9.643).

Για την περίπτωση που έχουμε παρατηρήσεις κατά ζεύγη επιλέγουμε:



Εικόνα 1.5.4: Επιλογή ελέγχου παρατηρήσεις κατά ζεύγη

Στη συνέχεια καθορίζουμε το ζεύγος των μεταβλητών και εφαρμόζουμε τον έλεγχο.

Η ανάλυση των αποτελεσμάτων γίνεται με τον ίδιο τρόπο.

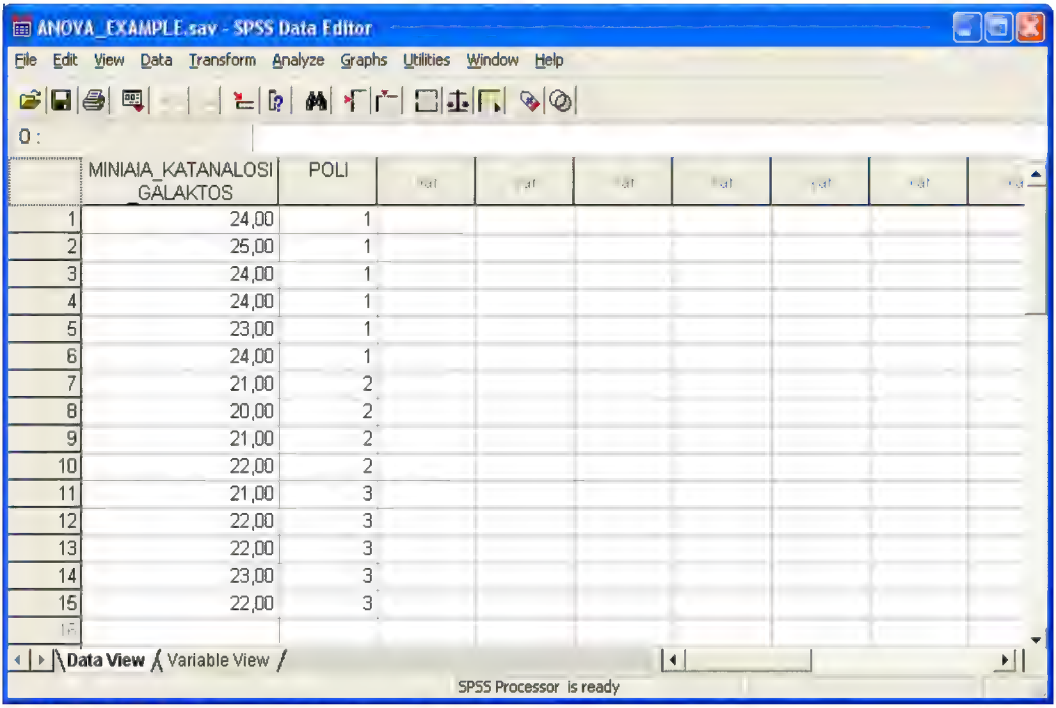
1.6 Πλήρως Τυχαιοποιημένος Σχεδιασμός με χρήση του SPSS

Σε αυτήν την ενότητα θα δούμε πώς εφαρμόζουμε όλα τα παραπάνω με τη χρήση του SPSS. Για να κατανοήσουμε πιο εύκολα τη μεθοδολογία θα χρησιμοποιήσουμε το προηγούμενο παράδειγμα. Έχουμε λοιπόν τον πίνακα δεδομένων:

	ΠΟΛΕΙΣ		
#Οικογενειών	Α	Β	Γ
1	24	21	21
2	25	20	22
3	24	21	22
4	24	22	23
5	23		22
6	24		

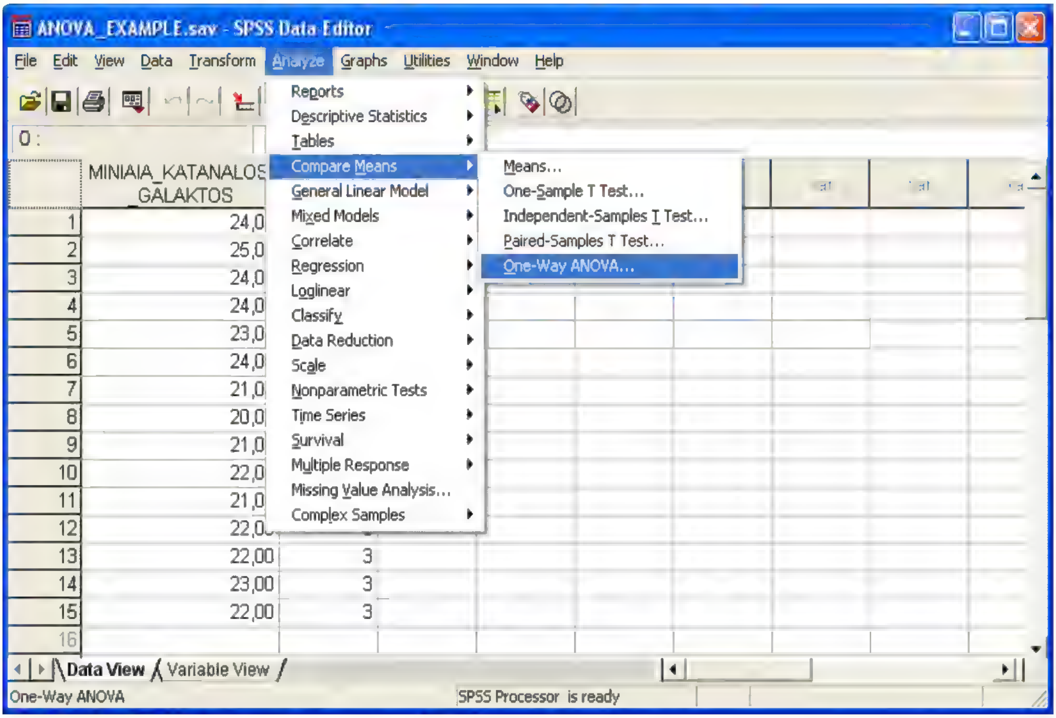
Πίνακας 1.6.1: Μηνιαία κατανάλωση γάλακτος

Πρώτος μας στόχος είναι να εισαγάγουμε τα δεδομένα στην καρτέλα δεδομένων. Εισάγουμε όλα τα δεδομένα στην πρώτη στήλη και την ονομάζουμε π.χ **MINIAIA_KATANALOSI_GALAKTOS**. Αυτή αποτελεί την εξαρτημένη μας μεταβλητή. Στη διπλανή στήλη θα πρέπει να εισαγάγουμε τις αντίστοιχες πόλεις από τις οποίες πήραμε το δείγμα. Επειδή, ωστόσο, η μέθοδος αυτή χρειάζεται αριθμητικά δεδομένα σε όλες τις τιμές που αφορούν την πόλη Α αντιστοιχούμε την τιμή 1. Ομοίως στις τιμές των δεδομένων της πόλης Β αντιστοιχούμε τον αριθμό 2 και σε αυτές της πόλης Γ τον αριθμό 3. Αυτό θα συνεχιζόταν με τον ίδιο τρόπο αν είχαμε και άλλα δείγματα (sample). Τη νέα αυτή μεταβλητή την ονομάζουμε **POLI**. Επομένως, έχουμε τα δεδομένα στη μορφή:



Εικόνα 1.6.1: Εισαγωγή δεδομένων στο SPSS

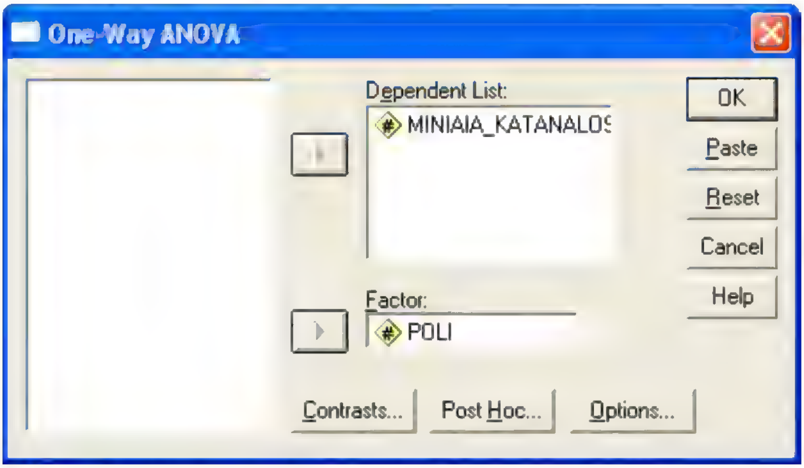
Το επόμενο βήμα είναι να εφαρμόσουμε τη μέθοδο. Ακολουθούμε αναλυτικά τα βήματα:



Εικόνα 1.6.2: Βήματα εφαρμογής ANOVA

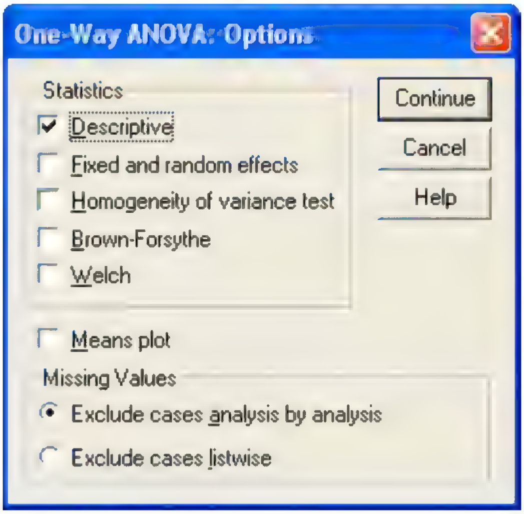
Στο κεντρικό menu της μεθόδου θα πρέπει να καθορίσουμε ποια μεταβλητή είναι η εξαρτημένη και ως προς ποια μεταβλητή θέλουμε να κάνουμε τον έλεγχο περί ισότητας των μέσων (παράγοντας).

Στο παράδειγμα μας θέλουμε να ελέγξουμε την υπόθεση περί ισότητας των μέσων της μηνιαίας κατανάλωσης γάλακτος (εξαρτημένη μεταβλητή) ως προς τις τρεις διαφορετικές πόλεις (παράγοντας). Επομένως έχουμε:



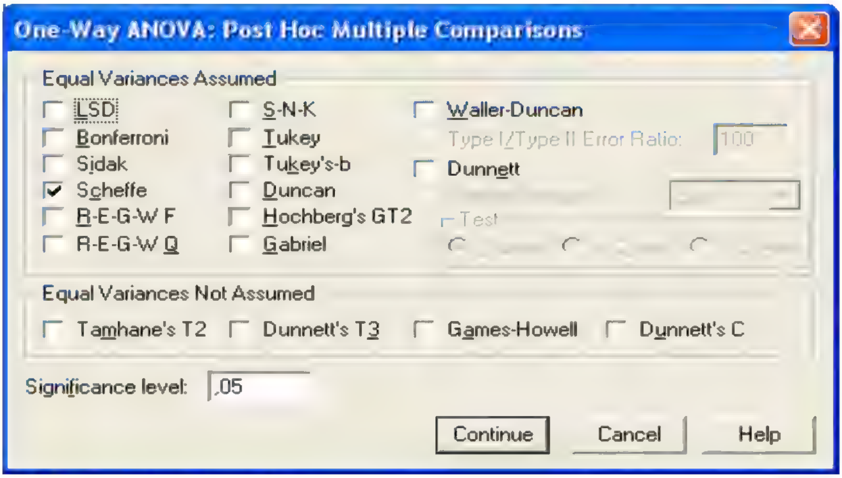
Εικόνα 1.6.3: Επιλογή εξαρτημένης επιλογής και παράγοντα

Στην καρτέλα **Options** μπορούμε να επιλέξουμε κάποια περιγραφικά μέτρα **(Descriptive)**. Αναλυτικά:



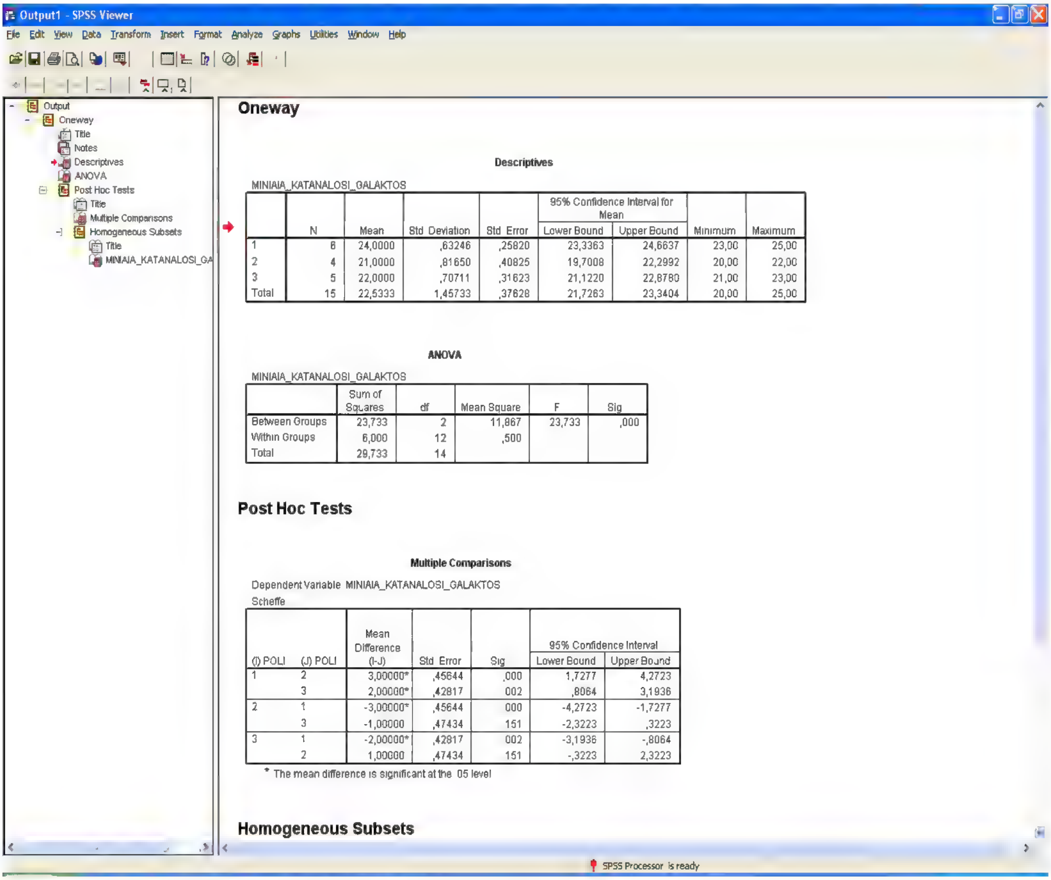
Εικόνα 1.6.4: Καρτέλα Options

Επίσης, στην καρτέλα Post Hoc έχουμε τη δυνατότητα να επιλέξουμε κάποιο test που εντοπίζει ανάμεσα σε ποιους μέσους υπάρχει διαφορά αν και εφόσον απορριφθεί η μηδενική υπόθεση. Έστω ότι επιλέγουμε τον έλεγχο του Scheffe.



Εικόνα 1.6.5: Καρτέλα Post Hoc

Εκτελώντας λοιπόν την εντολή παίρνουμε τα παρακάτω αποτελέσματα:



Oneway

Descriptives

MINIAIA_KATANALOSI_GALAKTOS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
1	6	24,0000	,63246	,25820	23,3363	24,6637	23,00	25,00
2	4	21,0000	,81650	,40825	19,7008	22,2992	20,00	22,00
3	5	22,0000	,70711	,31623	21,1220	22,8780	21,00	23,00
Total	15	22,5333	1,45733	,37628	21,7263	23,3404	20,00	25,00

ANOVA

MINIAIA_KATANALOSI_GALAKTOS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	23,733	2	11,867	23,733	,000
Within Groups	6,000	12	,500		
Total	29,733	14			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: MINIAIA_KATANALOSI_GALAKTOS
Scheffe

(I) POLI	(J) POLI	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	2	3,00000*	,45644	,000	1,7277	4,2723
	3	2,00000*	,42817	,002	,8064	3,1936
2	1	-3,00000*	,45644	,000	-4,2723	-1,7277
	3	-1,00000	,47434	,151	-2,3223	,3223
3	1	-2,00000*	,42817	,002	-3,1936	-,8064
	2	1,00000	,47434	,151	-,3223	2,3223

* The mean difference is significant at the .05 level.

Homogeneous Subsets

Εικόνα 1.6.6: One Way ANOVA

Ο πρώτος πίνακας αφορά κάποια περιγραφικά μέτρα.

Descriptives

MINIAIA_KATANALOSI_GALAKTOS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	6	24,0000	,63246	,25820	23,3363	24,6637	23,00	25,00
2	4	21,0000	,81650	,40825	19,7008	22,2992	20,00	22,00
3	5	22,0000	,70711	,31623	21,1220	22,8780	21,00	23,00
Total	15	22,5333	1,45733	,37628	21,7263	23,3404	20,00	25,00

Εικόνα 1.6.7: Περιγραφικά Μέτρα

Βλέπουμε λοιπόν ότι έχουμε τις 3 πόλεις. Φαίνεται το πλήθος του κάθε δείγματος ανά πόλη, η μέση τιμή στη μηνιαία κατανάλωση γάλακτος, η τυπική απόκλιση ένα 95% διάστημα εμπιστοσύνης, καθώς επίσης και η μικρότερη και η μέγιστη τιμή, πάλι ανά πόλη.

Ο επόμενος πίνακας είναι και ο βασικός και αποτυπώνει τον πίνακα Ανάλυσης Διακύμανσης

ANOVA

MINIAIA_KATANALOSI_GALAKTOS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	23,733	2	11,867	23,733	,000
Within Groups	6,000	12	,500		
Total	29,733	14			

Εικόνα 1.6.8: Πίνακας Ανάλυσης Διακύμανσης

Προφανώς, ο πίνακας που παίρνουμε ταυτίζεται με αυτόν που είχαμε κατασκευάσει παραπάνω. Εκτός της τιμής F (23.733) το SPSS μας δίνει και την τιμή του επίπεδου σημαντικότητας p που αντιστοιχεί σε αυτήν και είναι σχεδόν 0. Έτσι έχουμε:

$\mathbf{H_0 : m_1 = m_2 = \dots = m_K}$

$\mathbf{H_1}$ **: τουλάχιστον δύο μέσοι διαφέρουν**

Απορρίπτουμε τη μηδενική υπόθεση H_0 όταν το p έχει τιμή μικρότερη ή ίση με την τιμή του επιπέδου σημαντικότητας α που έχει γίνει η μέθοδος. Το α έχει οριστεί στο 0.05 και επειδή $p<0.05$ απορρίπτουμε τη μηδενική υπόθεση περί ισότητας των μέσων. Επομένως, έχουμε καταλήξει στο συμπέρασμα ότι τουλάχιστον δύο μέσες τιμές διαφέρουν. Το ερώτημα είναι ποιες.

Για αυτό το λόγο χρησιμοποιήσαμε την επιλογή των **Post Hoc** ελέγχων. Έχουμε:

Multiple Comparisons

Dependent Variable: MINIAIA_KATANALOSI_GALAKTOS
Scheffe

(I) POLI	(J) POLI	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	3,00000*	,45644	,000	1,7277	4,2723
	3	2,00000*	,42817	,002	,8064	3,1936
2	1	-3,00000*	,45644	,000	-4,2723	-1,7277
	3	-1,00000	,47434	,151	-2,3223	,3223
3	1	-2,00000*	,42817	,002	-3,1936	-,8064
	2	1,00000	,47434	,151	-,3223	2,3223

*. The mean difference is significant at the .05 level.

Εικόνα 1.6.9: Έλεγχος Sceffe για την ανίχνευση διαφορών στους μέσους

Όπως παρατηρούμε ο έλεγχος συγκρίνει τις μέσες τιμές ανά πόλη. Δηλαδή, πρώτα τη μέση τιμή της Α πόλης με αυτήν της Β, έπειτα της Α με της Γ και τέλος της Β με τη Γ. Σε κάθε μία περίπτωση διενεργείται ο έλεγχος:

$$H_0 : \mu_i = \mu_j$$

$$H_1 : \mu_i \neq \mu_j$$

--> Στην πρώτη περίπτωση (Α με Β) το p είναι μικρότερο του 0.05 οπότε απορρίπτεται η μηδενική υπόθεση. Οι μέσες τιμές της μηνιαίας κατανάλωσης γάλακτος μεταξύ των πόλεων Α και Β διαφέρουν στατιστικά σημαντικά.

--> Στη δεύτερη περίπτωση (Α με Γ) το p είναι μικρότερο του 0.05 (p=0.02) οπότε απορρίπτεται και εδώ η μηδενική υπόθεση. Άρα, οι μέσες τιμές της μηνιαίας κατανάλωσης γάλακτος μεταξύ των πόλεων Α και Γ διαφέρουν στατιστικά σημαντικά.

--> Στην τρίτη περίπτωση (Β με Γ) το p είναι μεγαλύτερο του 0.05 (p=0.151) οπότε δεν απορρίπτεται η μηδενική υπόθεση. Άρα, οι μέσες τιμές της μηνιαίας κατανάλωσης γάλακτος μεταξύ των πόλεων Β και Γ δεν διαφέρουν στατιστικά σημαντικά.

Μπορούμε βέβαια αντί για το test του Scheffe να εφαρμόσουμε οποιονδήποτε άλλο από τους ελέγχους που μας παρέχονται.

1.7 Έλεγχος Ανεξαρτησίας Μεταβλητών με χρήση του SPSS

Σε αυτήν την ενότητα θα δούμε με ποιο τρόπο χειριζόμαστε τα δεδομένα όταν θέλουμε να χρησιμοποιήσουμε το SPSS για να εφαρμόσουμε τους πίνακες συνάφειας. Έχουμε τον πίνακα των δεδομένων:

	Ψυχολογικό τεστ		
Τεστ Δεξιοτήτων	Εσωστρεφείς	Εξωστρεφείς	Σύνολο
Επιτυχόντες	13	73	86
Αποτυχόντες	17	57	74
Σύνολο	**30**	**130**	**160**

Πίνακας 1.7.1: Πίνακας δεδομένων

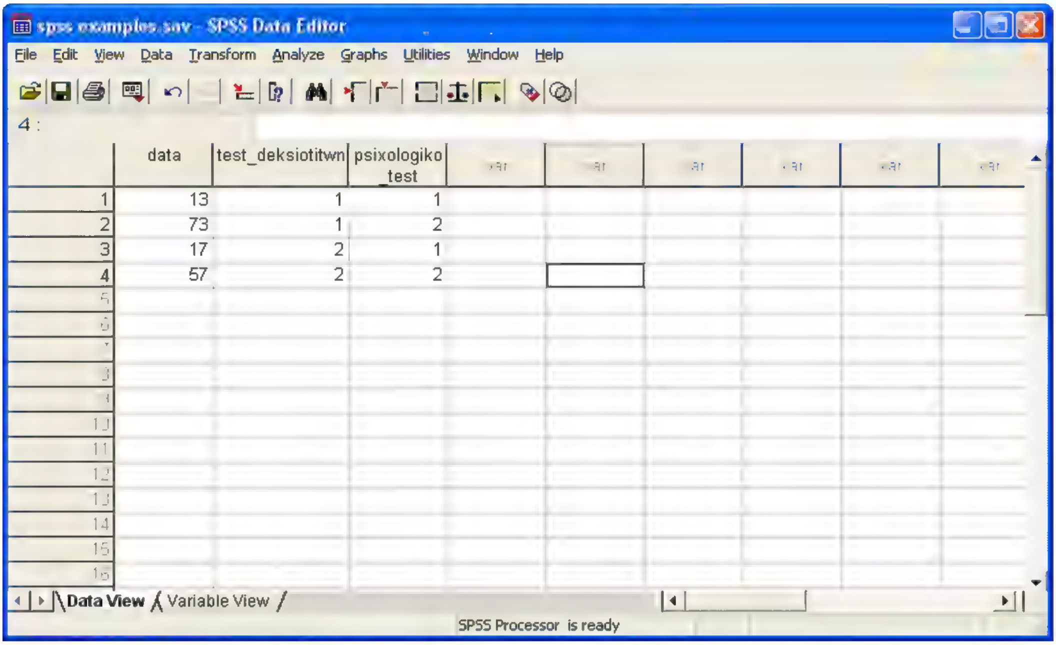
Αρχικά θα πρέπει να καταχωρήσουμε τα δεδομένα μας. Δημιουργούμε τρεις μεταβλητές.

□ Η πρώτη θα ονομάζεται data και θα περιλαμβάνει τα δεδομένα μας, δηλαδή τον αριθμό των ατόμων που ανήκουν σε κάθε κατηγορία.

□ Η δεύτερη θα ονομάζεται test_deksiotitwn και θα παίρνει τις τιμές 1 και 2. 1 αν τα δεδομένα αφορούν τους επιτυχόντες του τεστ δεξιοτήτων και 2 αν αφορούν τους αποτυχόντες.

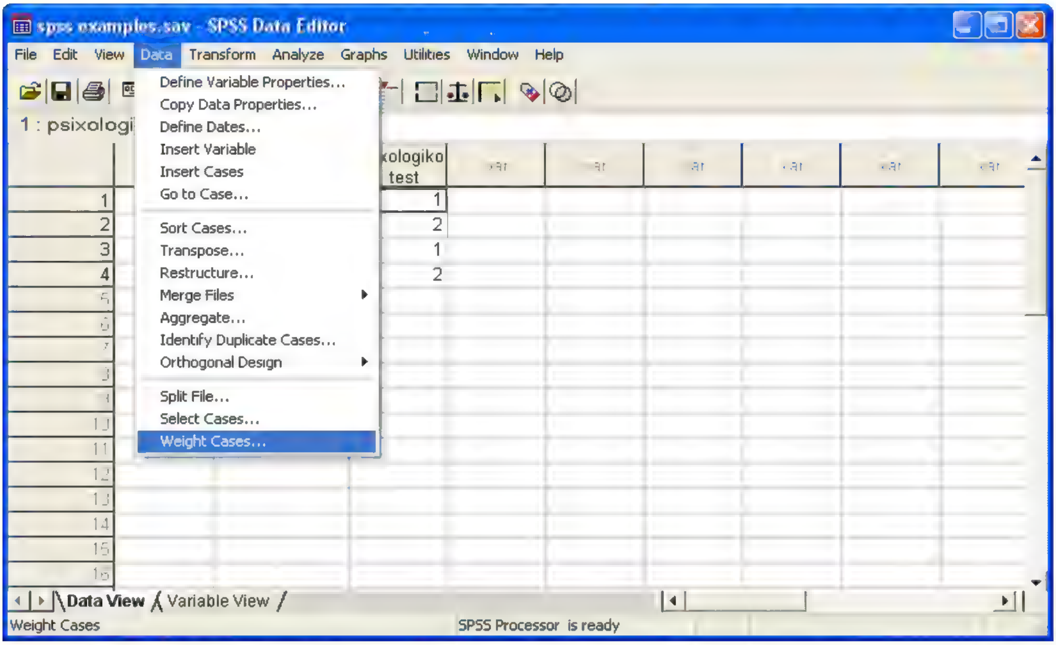
□ Η τρίτη μεταβλητή θα ονομάζεται psixologiko_test και θα παίρνει τις τιμές 1 εφόσον αφορά "εσωστρεφείς" πιλότους και 2 εφόσον αφορά «εξωστρεφείς».

Τα δεδομένα μας λοιπόν έχουν την παρακάτω μορφή:



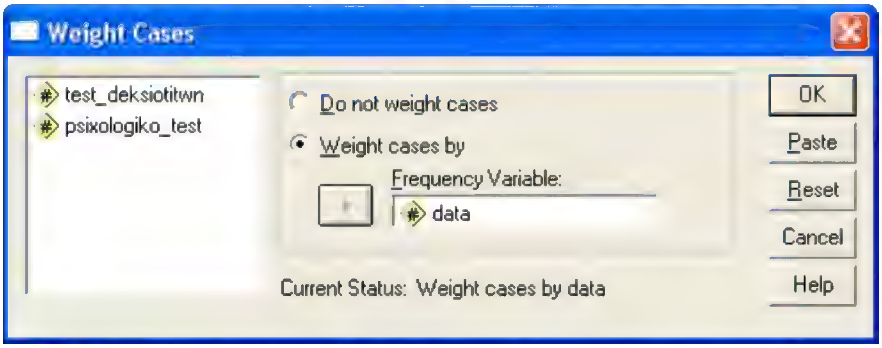
Εικόνα 1.7.1: Εισαγωγή δεδομένων στο SPSS

Είμαστε έτοιμοι λοιπόν να εφαρμόσουμε τη μέθοδο. Ακολουθούμε τα βήματα:



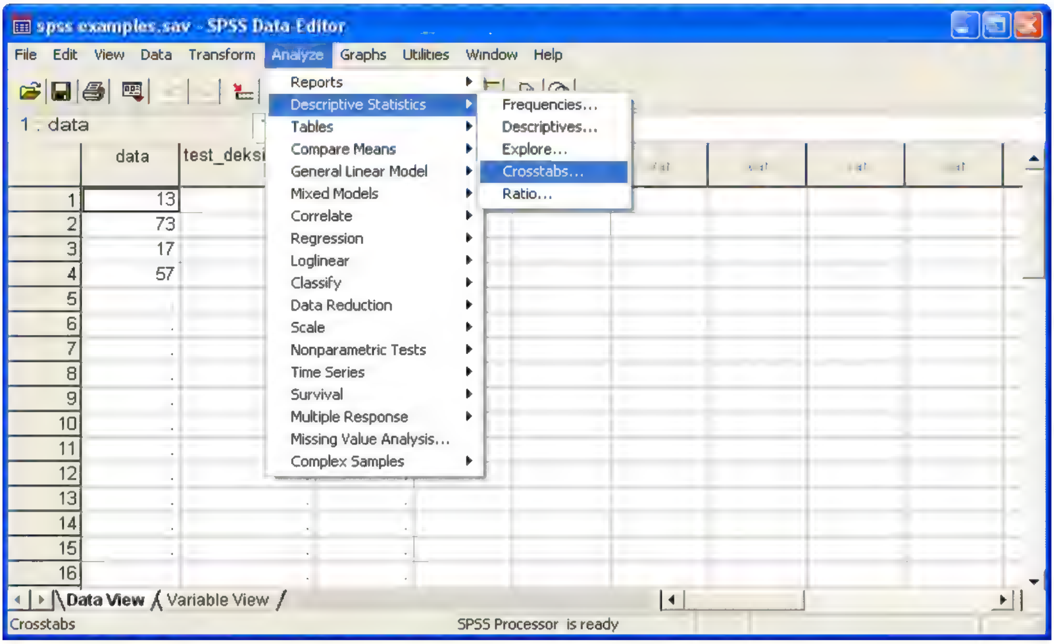
Εικόνα 1.7.2: 1ο Βήμα Ελέγχου Ανεξαρτησίας

Σε αυτό το βήμα καθορίζουμε ως προς ποια μεταβλητή σταθμίζουμε τα δεδομένα ώστε να εφαρμόσουμε τον έλεγχο ανεξαρτησίας. Έτσι:



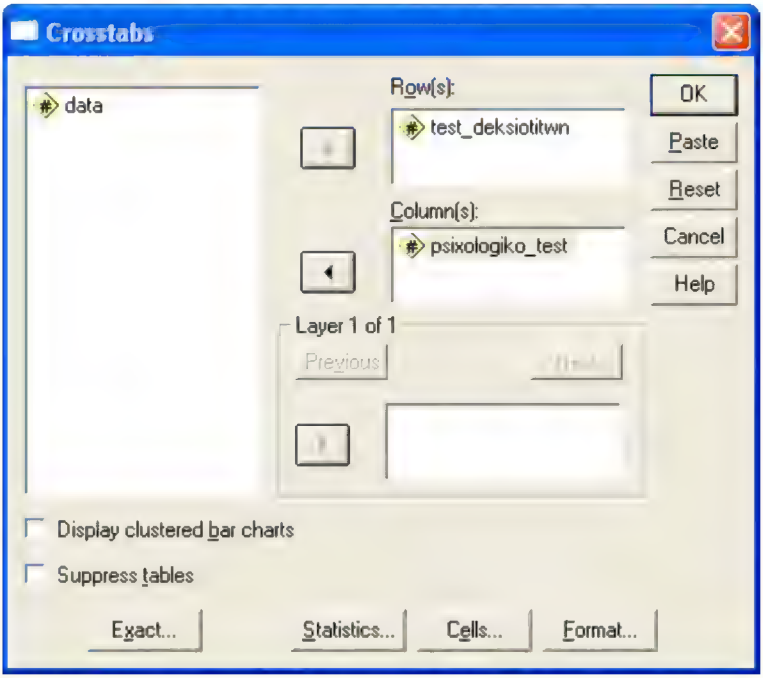
Εικόνα 1.7.3: Καρτέλα επιλογής σταθμισμένης μεταβλητής

Θέλουμε ο έλεγχος να γίνει ως προς τον αριθμό των ατόμων που ανήκουν σε κάθε συνδυασμό και για αυτό επιλέγουμε τη μεταβλητή **data** ως σταθμισμένη μεταβλητή. Είμαστε έτοιμοι πλέον να εφαρμόσουμε τη διαδικασία. Από το menu επιλέγουμε:



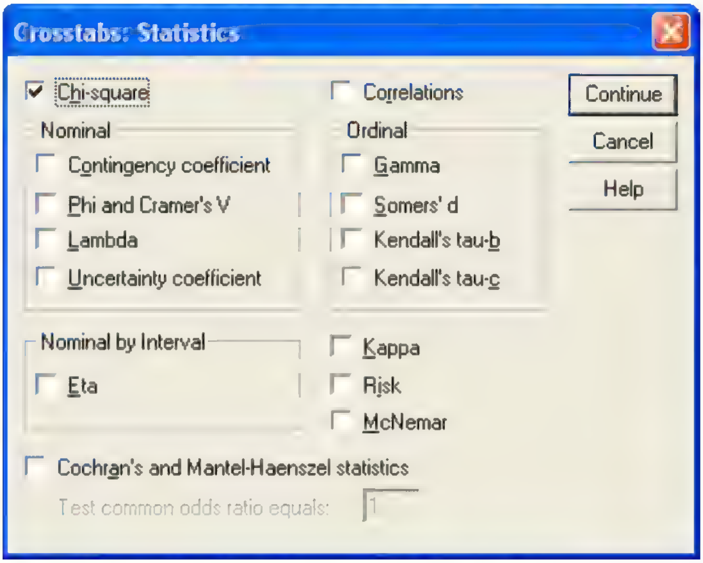
Εικόνα 1.7.4: Επιλογή Ελέγχου Συνάφειας

Στο κεντρικό menu καθορίζουμε ποια μεταβλητή θα ανήκει στη γραμμή και ποια στη στήλη του πίνακα που θέλουμε να δημιουργήσουμε. Έτσι:



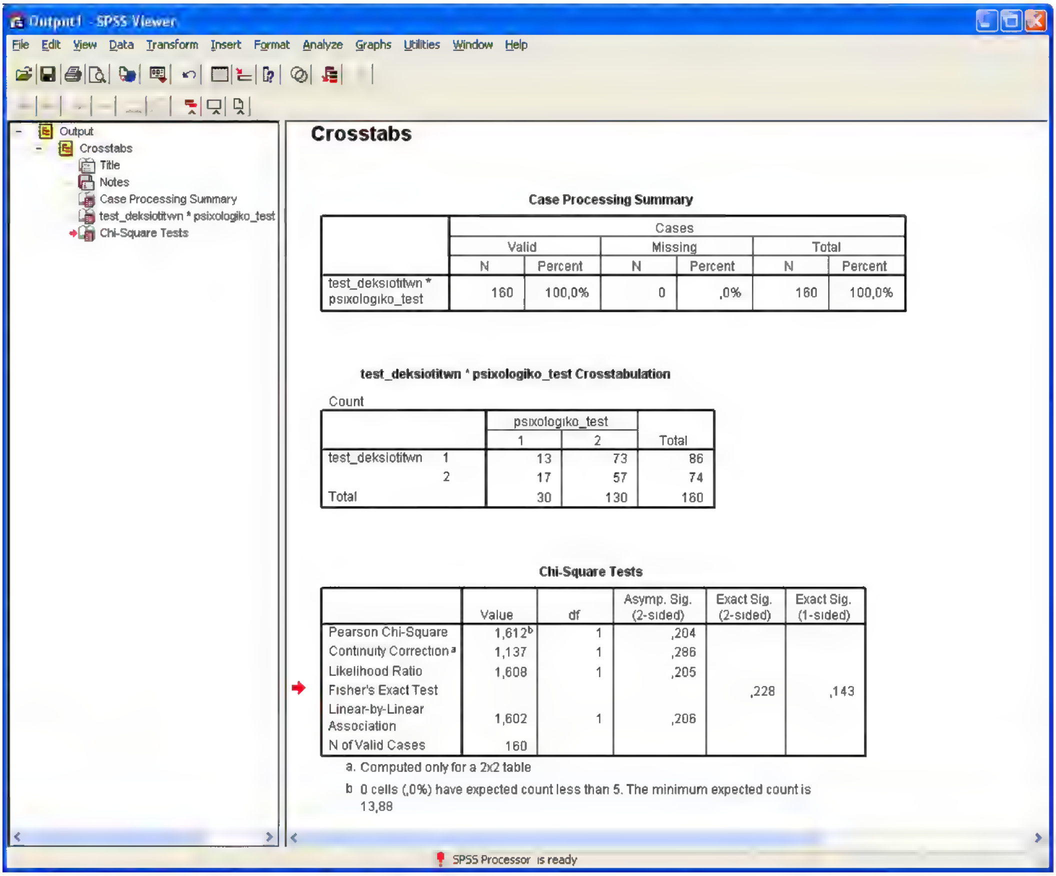
Εικόνα 1.7.5 : Καρτέλα καθορισμού μεταβλητών

Τέλος, από την καρτέλα **Statistics** επιλέγουμε τον έλεγχο ανεξαρτησίας χ^{σ} για να διαπιστώσουμε τι συμβαίνει μεταξύ των μεταβλητών.



Εικόνα 1.7.6: Καρτέλα επιλογής ελέγχου ανεξαρτησίας

Εφαρμόζοντας όλα τα παραπάνω παίρνουμε από το SPSS τα παρακάτω:



Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
test_deksiotitwn * psixologiko_test	160	100,0%	0	,0%	160	100,0%

test_deksiotitwn * psixologiko_test Crosstabulation

Count

		psixologiko_test		
		1	2	Total
test_deksiotitwn	1	13	73	86
	2	17	57	74
Total		30	130	160

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1,612[b]	1	,204		
Continuity Correction[a]	1,137	1	,286		
Likelihood Ratio	1,608	1	,205		
Fisher's Exact Test				,228	,143
Linear-by-Linear Association	1,602	1	,206		
N of Valid Cases	160				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 13,88

Εικόνα 1.7.7 : Πίνακας Συνάφειας 2Χ2

Ο έλεγχος που θέλαμε να κάνουμε είναι:

H_0: Η ικανότητα του πιλότου είναι ανεξάρτητη από τον τύπο της προσωπικότητάς του.
H_1: Η ικανότητα του πιλότου σχετίζεται με τον τύπο της προσωπικότητάς του.

Από τον πίνακα φαίνεται ότι το **p-value** έχει τιμή **0.204** και αφού αυτή η τιμή είναι μεγαλύτερη από 0.05 (το επίπεδο που πραγματοποιούμε τον έλεγχο) δεν απορρίπτουμε τη μηδενική υπόθεση. Επομένως, δε μπορούμε να θεωρήσουμε ότι η ικανότητα ενός πιλότου σχετίζεται με τον τύπο της προσωπικότητας του.

1.8 Εισαγωγή στους μη παραμετρικούς ελέγχους

Παράδειγμα

Θέλουμε να ελέγξουμε κατά πόσο η ημερήσια παραγωγή γάλακτος προσεγγίζεται από την κανονική κατανομή με μ, σ_2 παραμέτρους. Τα δεδομένα μας είναι:

Ημερήσια παραγωγή γάλακτος 40 ημερών			
16,93	14,62	15,79	13,20
16,12	18,74	13,32	16,40
18,79	15,04	18,08	16,32
18,04	13,25	16,56	20,55
13,98	18,05	16,16	14,20
18,79	13,98	12,39	16,08
17,81	15,99	13,63	13,76
18,36	18,79	17,32	17,54
13,00	12,43	14,12	16,75
16,58	13,29	15,25	18,23

Πίνακας 1.8.1: **Πίνακας δεδομένων**

Θέλουμε λοιπόν να κάνουμε τον έλεγχο:

$$H_0 : F_X(x) = F_{N(\mu,\sigma^2)}(x)$$
$$H_1 : F_X(x) \neq F_{N(\mu,\sigma^2)}(x)$$

Το πρώτο βήμα που πρέπει να κάνουμε είναι να εισαγάγουμε τα δεδομένα στο SPSS. Τα καταχωρούμε στη μεταβλητή ***x***. Έπειτα, αποφασίζουμε τον αριθμό των κλάσεων που θα χρησιμοποιήσουμε. Καταλήγουμε ότι θα ταξινομήσουμε τα δεδομένα μας σε 8 κλάσεις έτσι ώστε να ισχύει:

$$p_i^o = \text{P (η παρατήρηση } X_j \text{ να ανήκει στην κλάση i } | \ H_0) = 1/8 = 0.125$$

Επομένως τα ποσοστιαία σημεία όλων των κλάσεων είναι:
0.125, 0.25, 0,375, 0.5, 0.625, 0.75, 0.875

Καταχωρούμε τα δεδομένα στη μεταβλητή, έστω ***p***.

Από την καρτέλα **Frequencies** βρίσκουμε τα μ και **σ^2**. Έχουμε:

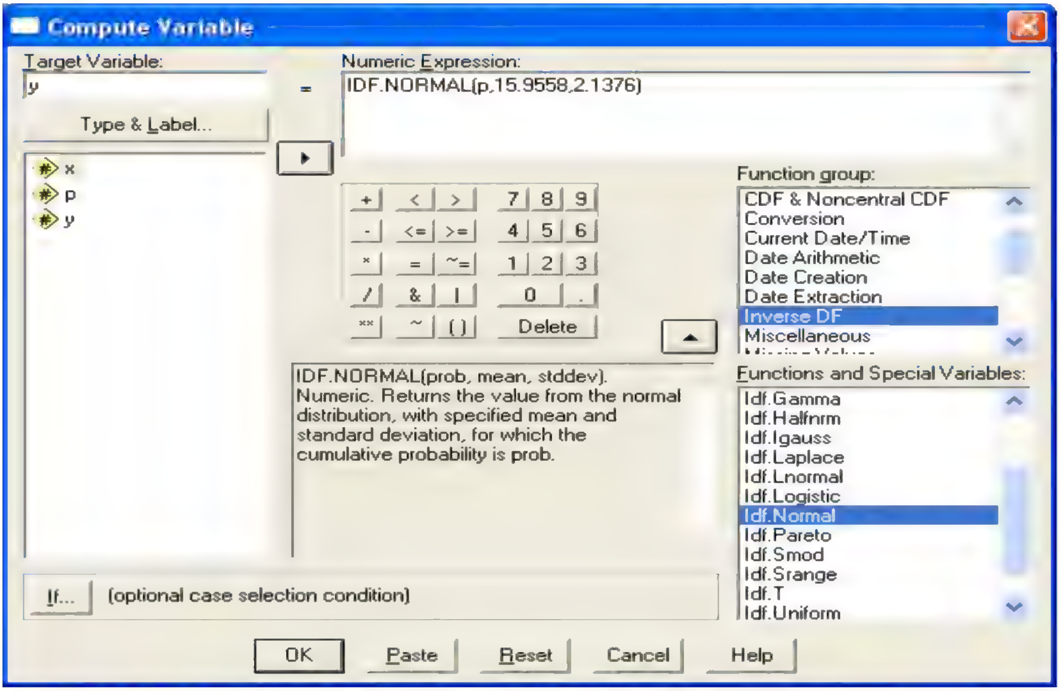
Statistics

x

N	Valid	40
	Missing	0
Mean		15,9558
Std. Deviation		2,13759

Εικόνα 1.8.1: Περιγραφικά στοιχεία

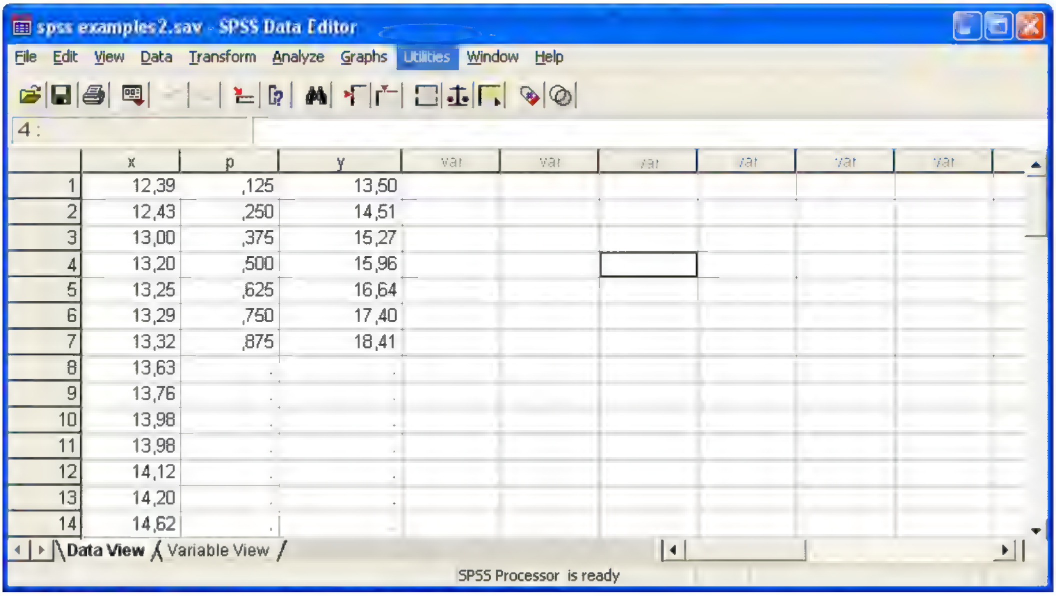
Το επόμενο βήμα είναι να δημιουργήσουμε τα όρια για την κάθε κλάση μας. Θα τα τοποθετήσουμε στη μεταβλητή ***y***. Στην καρτέλα **Compute Variable** κάνουμε το εξής:



Εικόνα 1.8.2 : Δημιουργία ορίων στις κλάσεις μας

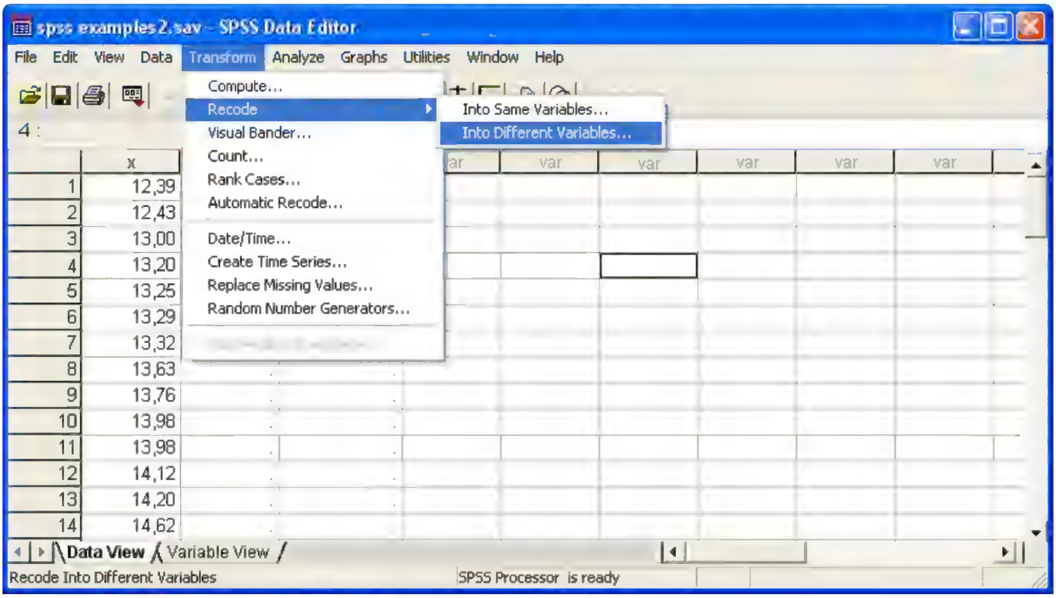
Στο **πεδίο –> IDF.NORMAL(p,15.9558,2.1376)** καθορίζουμε τη μεταβλητή που περιέχει τα ποσοστιαία σημεία των κλάσεων μας (**p**), τη μέση τιμή (**μ=15,9558**) και τη διακύμανση (**σ^2 = 2,1376**) των δεδομένων μας.

Επομένως, το κεντρικό menu έχει την ακόλουθη μορφή:



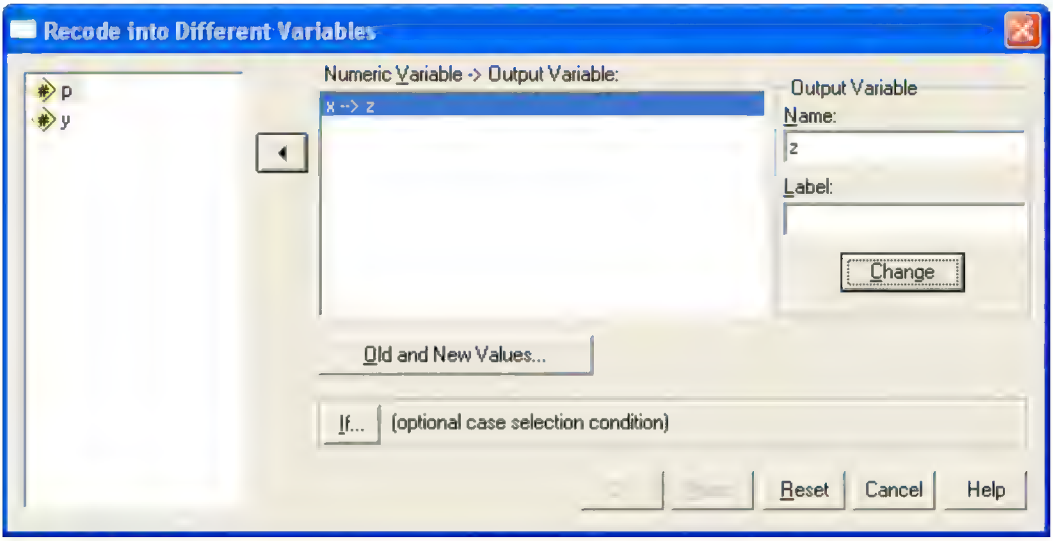
Εικόνα 1.8.3: Κεντρικό menu SPSS

Η μεταβλητή ***y*** περιλαμβάνει τα όρια των 8 κλάσεων μας. Στη συνέχεια θα κατασκευάσουμε άλλη μία μεταβλητή (έστω ***z***) που ουσιαστικά θα είναι η κάθε κλάση. Έτσι, επιλέγουμε:



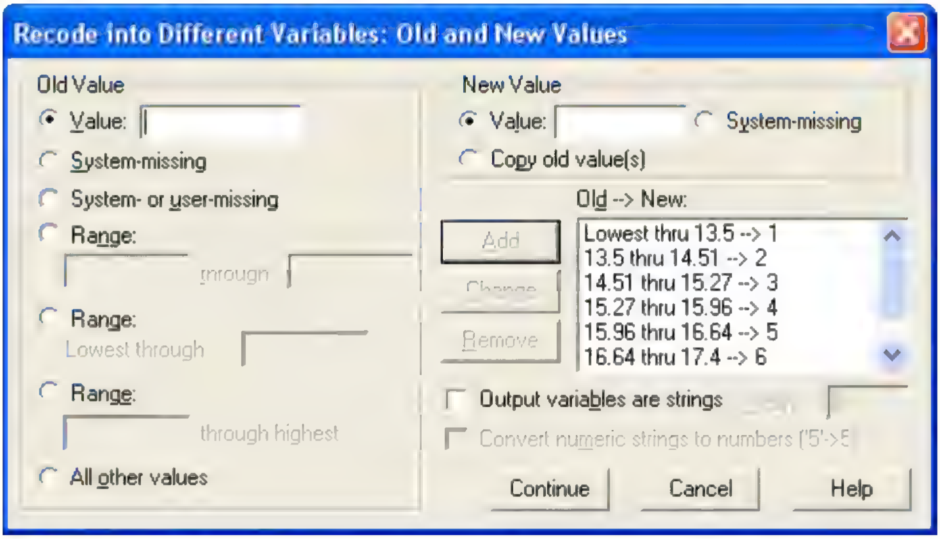
Εικόνα 1.8.4: Δημιουργία κλάσεων

Έπειτα καθορίζουμε ποια μεταβλητή θα επανακαθορίζουμε (***x***) και πώς θέλουμε να ονομάζεται (καρτέλα **output variable**)



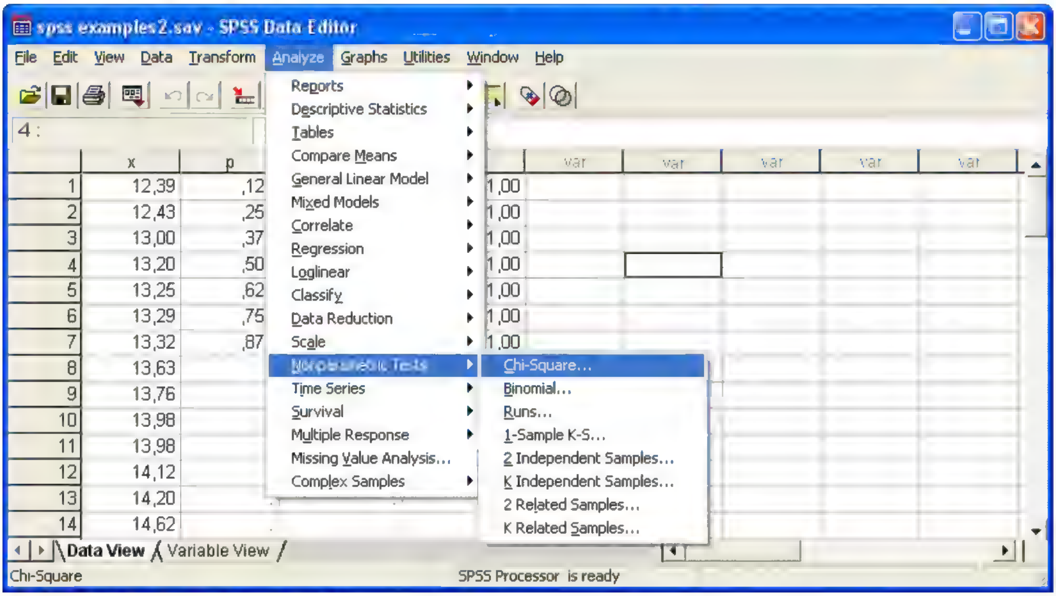
Εικόνα 1.8.5: Δημιουργία μεταβλητής με συγκεκριμένα όρια

Στη συνέχεια εισάγουμε τα όρια μας για την κάθε κλάση από την καρτέλα **Old** and **New Values**. Στην καρτέλα **Old Value** εισάγουμε το εύρος των τιμών και στην καρτέλα **New Value** την καινούργια τιμή. Για την πρώτη κλάση (τιμή 13,5 σύμφωνα με τη μεταβλητή ***y***) επιλέγουμε την καρτέλα **Lowest through** και για την τελευταία κλάση (τιμή 18,41 σύμφωνα με τη μεταβλητή ***y***) την καρτέλα **through highest**. Έχουμε λοιπόν:



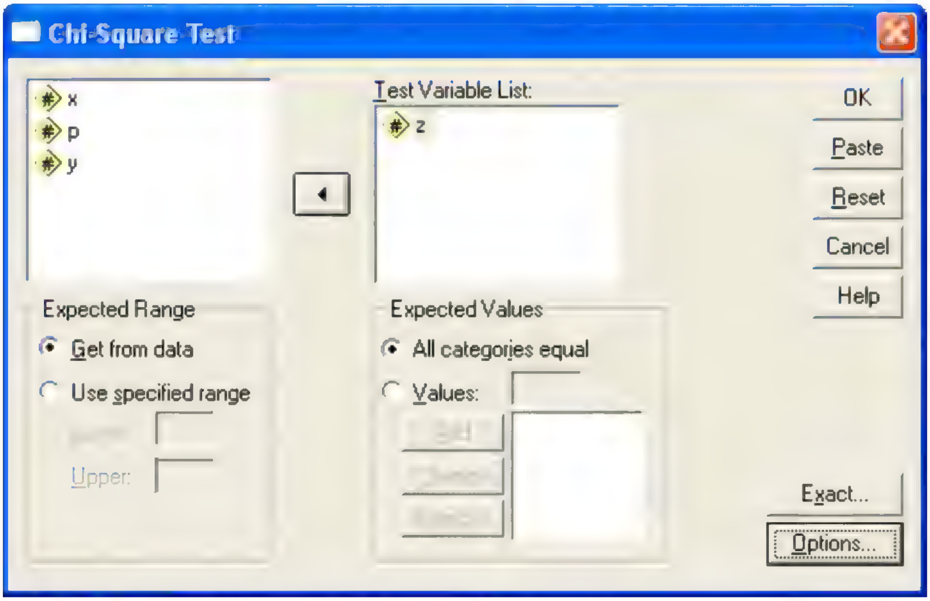
Εικόνα 1.8.6: Δημιουργία ορίων

Είμαστε έτοιμοι πλέον να εφαρμόσουμε τον έλεγχο μας. Επιλέγουμε στο menu:



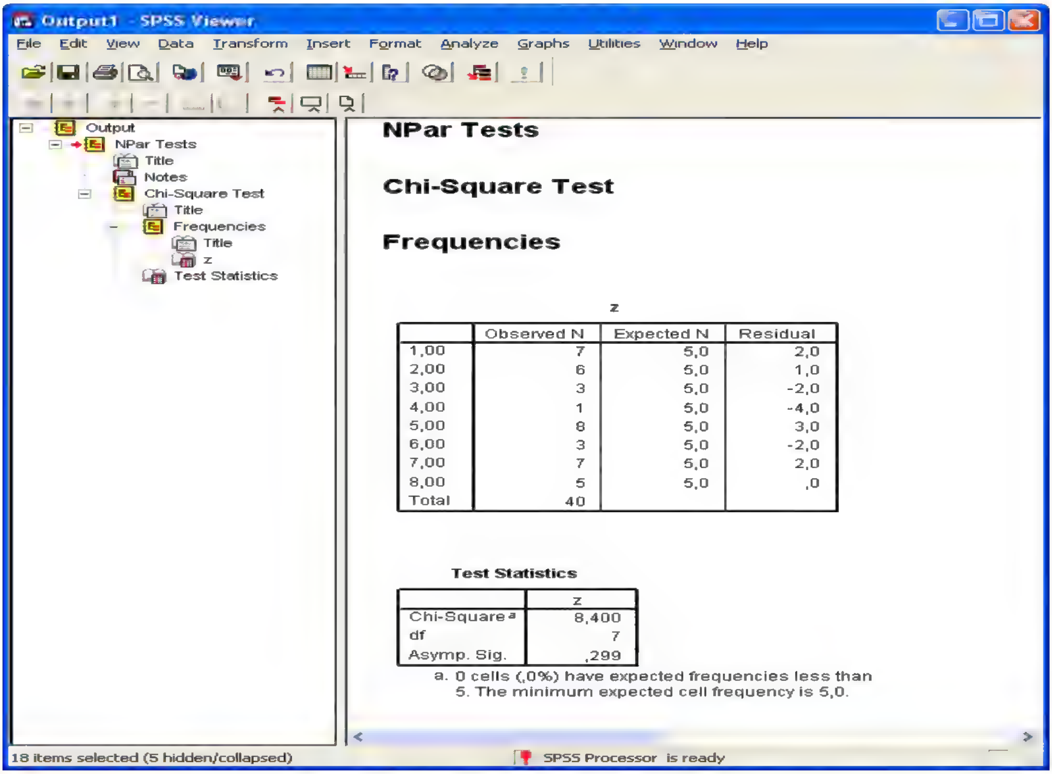
Εικόνα 1.8.7: Επιλογή ελέγχου Chi-Square

Καθορίζουμε σε ποια μεταβλητή θέλουμε να γίνει ο έλεγχος. Επιλέγουμε την καρτέλα **All categories equal**, αφού οι κλάσεις είναι ισοπίθανες. Έχουμε λοιπόν:



Εικόνα 1.8.8: Κεντρικό menu ελέγχου Chi-Square

Εφαρμόζω τον έλεγχο και παίρνω το εξής output.



z

	Observed N	Expected N	Residual
1,00	7	5,0	2,0
2,00	6	5,0	1,0
3,00	3	5,0	-2,0
4,00	1	5,0	-4,0
5,00	8	5,0	3,0
6,00	3	5,0	-2,0
7,00	7	5,0	2,0
8,00	5	5,0	,0
Total	40		

Test Statistics

	z
Chi-Square[a]	8,400
df	7
Asymp. Sig.	,299

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 5,0.

Εικόνα 1.8.9: Αποτελέσματα ελέγχου Chi-Square

Βλέπουμε την κατανομή των καινούργιων κλάσεων, καθώς επίσης και τα αποτελέσματα του ελέγχου στον πίνακα. Επειδή το p-value=0.299 έχει τιμή μεγαλύτερη από 0.05 δεν απορρίπτουμε τη μηδενική υπόθεση και επομένως θεωρούμε ότι όντως τα δεδομένα προέρχονται από μία κανονική κατανομή.

1.9 Runs Test

Ο επόμενος μη παραμετρικός έλεγχος με τον οποίο θα ασχοληθούμε είναι ο έλεγχος των ροών (runs test). Ο έλεγχος αυτός βασίζεται στον αριθμό των ροών που εμφανίζονται σε δείγμα παρατηρήσεων. Ως ροή εννοούμε τη διαδοχή όμοιων συμβόλων (π.χ. +, -) με άλλα σύμβολα. Για παράδειγμα:

```
++ ---- +++ - + ----- +++ -
 1   2    3  4 5  6     7  8
```

Έχουμε 8 ροές (8 φορές εναλλάσσονται τα σύμβολα). Όσο μικρότερος είναι ο αριθμός των ροών τόσο πιο ισχυρές ενδείξεις έχουμε ότι τα δεδομένα μας δεν εμφανίζονται με τυχαία σειρά, αλλά, υπάρχει κάποιου είδους τάση στη σειρά καταγραφής τους. Αντίθετα, μεγάλος αριθμός ροών ενισχύει τη θεωρία περί τυχαιότητας στην εμφάνιση των δεδομένων μας. Ο έλεγχος λοιπόν που κάνουμε είναι της μορφής:

H_0: Τα σύμβολα εμφανίζονται με τυχαία σειρά και επομένως τα δεδομένα μας κατανέμονται τυχαία.
H_1 : Η σειρά εμφάνισης των συμβόλων δεν είναι τυχαία και επομένως θεωρούμε ότι υπάρχει κάποιου είδους τάση στα δεδομένα μας.

Ο έλεγχος των υποθέσεων θα γίνει με τη στατιστική συνάρτηση Τ για την οποία ισχύει ότι αν $|T1| > z_{1-\alpha/2}$ τότε απορρίπτουμε τη μηδενική υπόθεση. Για να χρησιμοποιήσουμε την προηγούμενη κατανομή θα πρέπει το μέγεθος ενός από τα δύο σύμβολα να είναι μεγαλύτερο από 20. Σε διαφορετική περίπτωση χρησιμοποιούμε το $|T1| > w_{1-\alpha/2}$ όπου w_α είναι το α ποσοστιαίο σημείο της κατανομής της στατιστικής συνάρτησης Τ (ΠΑΡΑΡΤΗΜΑ). Ας δούμε ένα παράδειγμα.

Παράδειγμα

Χρησιμοποιούμε το ίδιο παράδειγμα της προηγούμενης ενότητας με την ημερήσια παραγωγή γάλακτος. Ο πίνακας των δεδομένων είναι:

Ημερήσια παραγωγή γάλακτος 40 ημερών			
16,93	14,62	15,79	13,20
16,12	18,74	13,32	16,40
18,79	15,04	18,08	16,32
18,04	13,25	16,56	20,55
13,98	18,05	16,16	14,20
18,79	13,98	12,39	16,08
17,81	15,99	13,63	13,76
18,36	18,79	17,32	17,54
13,00	12,43	14,12	16,75
16,58	13,29	15,25	18,23

Πίνακας 1.9.1: **Πίνακας δεδομένων**

Για να μπορέσουμε να πραγματοποιήσουμε τον έλεγχο χρειαζόμαστε μια παράμετρο ως προς την οποία θα συγκρίνουμε κάθε τιμή και θα αντιστοιχούμε ένα + αν είναι μεγαλύτερη ή ένα – αν είναι μικρότερη. Μπορούμε να χρησιμοποιήσουμε είτε τη μέση τιμή είτε τη διάμεσο. Έστω ότι χρησιμοποιούμε τη διάμεσο. Για τα παραπάνω δεδομένα αυτή είναι **16.14** και επομένως ο πίνακας γίνεται:

Ημερήσια παραγωγή γάλακτος 40 ημερών							
16,93	+	14,62	-	15,79	-	13,20	-
16,12	-	18,74	+	13,32	-	16,40	+
18,79	+	15,04	-	18,08	+	16,32	+
18,04	+	13,25	-	16,56	+	20,55	+
13,98	-	18,05	+	16,16	+	14,20	-
18,79	+	13,98	-	12,39	-	16,08	-
17,81	+	15,99	-	13,63	-	13,76	-
18,36	+	18,79	+	17,32	+	17,54	+
13,00	-	12,43	-	14,12	-	16,75	+
16,58	+	13,29	-	15,25	-	18,23	+

Πίνακας 1.9.2: Πίνακας δεδομένων

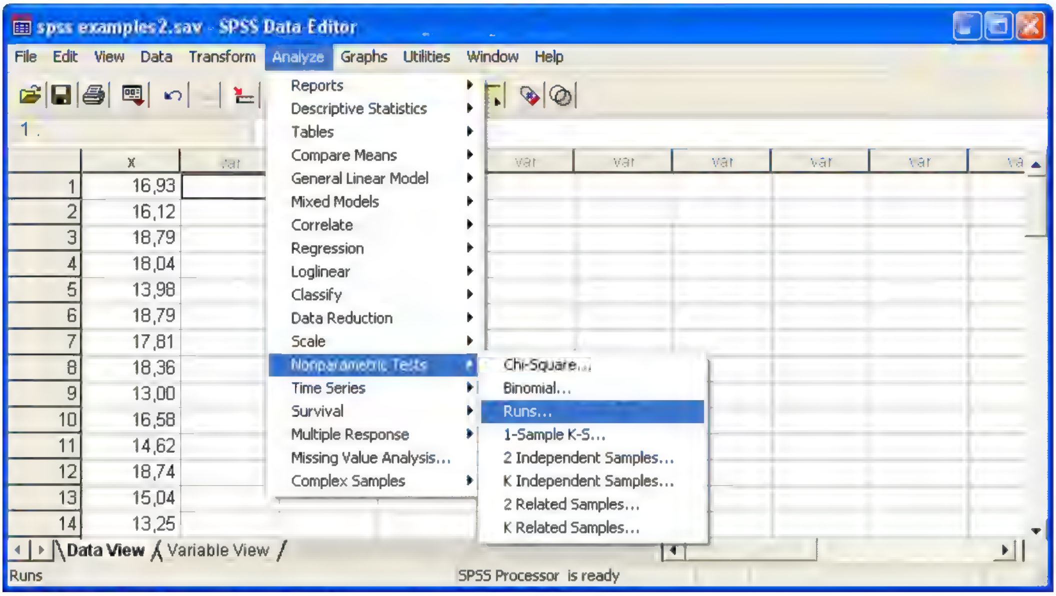
Μετρώντας τις εναλλαγές των συμβόλων καταλήγουμε σε ένα σύνολο T = 21 ροών. Το άθροισμα των + συμβόλων είναι 20 και των – 20 επομένως χρησιμοποιούμε την w κατανομή. Σε επίπεδο σημαντικότητας α = 0.05 έχουμε ότι (βλ. Παράρτημα):

$\mathbf{T \leq w_{a/2} = w_{0.025} = 14}$ και

$\mathbf{T \geq w_{1-a/2} = w_{0.075} = 28}$

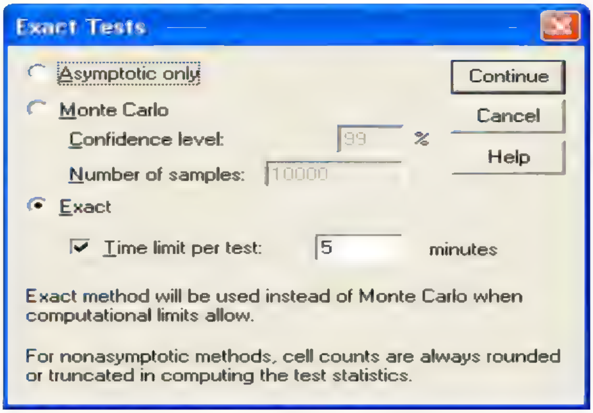
Εμείς έχουμε **T = 21** και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση περί τυχαιότητας των δεδομένων μας.

Έχουμε ήδη εισαγάγει τα δεδομένα μας. Επιλέγουμε:



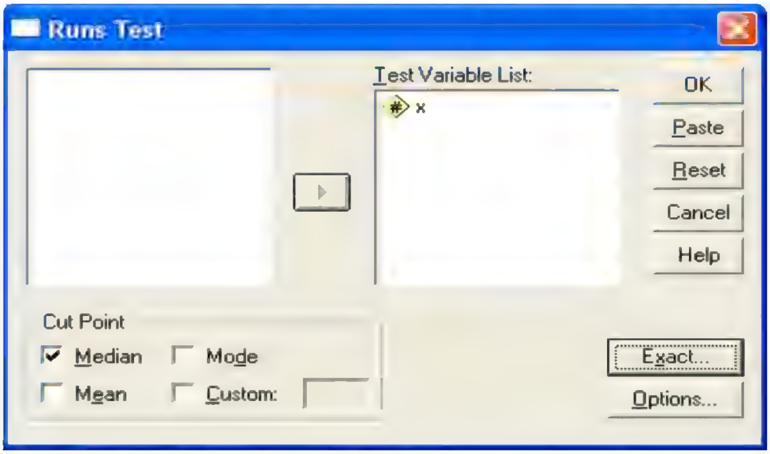
Εικόνα 1.9.1: Επιλογή ελέγχου Runs Test

Στην καρτέλα **Exact Tests** επιλέγουμε:



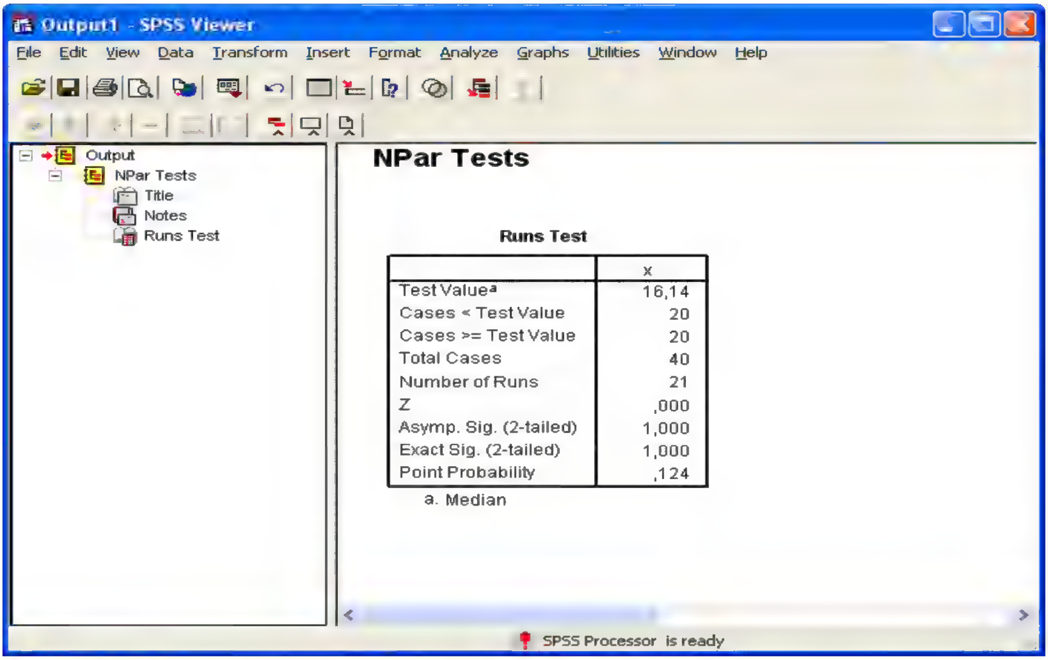
Εικόνα 1.9.2: Επιλογή Exact Runs Test

Κατόπιν επιλέγουμε τη μεταβλητή που θέλουμε να ελέγξουμε καθώς επίσης και ως προς ποια παράμετρο θα συγκρίνουμε την κάθε μας τιμή (καρτέλα **Cut Point)**.



Εικόνα 1.9.3: Κεντρικό menu Runs Test

Το αποτέλεσμα που παίρνουμε είναι το εξής:



Runs Test

	x
Test Value[a]	16,14
Cases < Test Value	20
Cases >= Test Value	20
Total Cases	40
Number of Runs	21
Z	,000
Asymp. Sig. (2-tailed)	1,000
Exact Sig. (2-tailed)	1,000
Point Probability	,124

a. Median

Εικόνα 1.9.4: Αποτελέσματα ελέγχου Runs Test

Το p-value έχει τιμή μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση. Άρα δε θεωρούμε ότι υπάρχει συγκεκριμένη τάση στη σειρά καταγραφής των δεδομένων μας.

1.10 Kolmogorov-Smirnov ενός δείγματος

Ο τελευταίος μη παραμετρικός έλεγχος με τον οποίο θα ασχοληθούμε είναι ο έλεγχος Kolmogorov-Smirnov για την περίπτωση ενός δείγματος. Χρησιμοποιείται για να ελέγξουμε την υπόθεση κατά πόσο τα δεδομένα μας προσεγγίζονται ικανοποιητικά από μια συγκεκριμένη κατανομή.

Η διαφορά του από τον Chi-Square είναι ότι ο Kolmogorov-Smirnov εφαρμόζεται και για διατεταγμένα δεδομένα καθώς επίσης και ότι δεν χρησιμοποιεί τη συνάρτηση πυκνότητας πιθανότητας αλλά, την αθροιστική συνάρτηση κατανομής $F(x) = P(X \leq x)$. Το πρόβλημα που έχει αυτός ο έλεγχος είναι ότι είναι πιο αποτελεσματικός σε περιπτώσεις που οι παράμετροι της κατανομής που θέλουμε να ελέγξουμε δεν είναι άγνωστοι, δε χρειάζεται δηλαδή να εκτιμηθούν. Σε διαφορετική περίπτωση ο Kolmogorov-Smirnov είναι πιο συντηρητικός.

Παράδειγμα

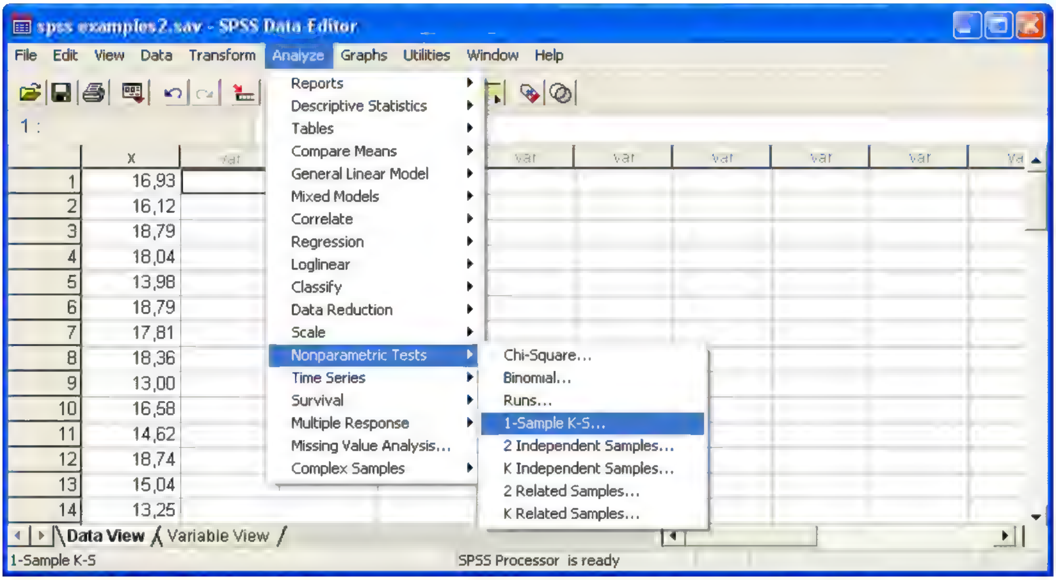
Χρησιμοποιούμε και πάλι το παράδειγμα με την ημερήσια παραγωγή γάλακτος. Ο πίνακας των δεδομένων είναι:

Ημερήσια παραγωγή γάλακτος 40 ημερών			
16,93	14,62	15,79	13,20
16,12	18,74	13,32	16,40
18,79	15,04	18,08	16,32
18,04	13,25	16,56	20,55
13,98	18,05	16,16	14,20
18,79	13,98	12,39	16,08
17,81	15,99	13,63	13,76
18,36	18,79	17,32	17,54
13,00	12,43	14,12	16,75
16,58	13,29	15,25	18,23

Πίνακας 1.10.1: Πίνακας δεδομένων

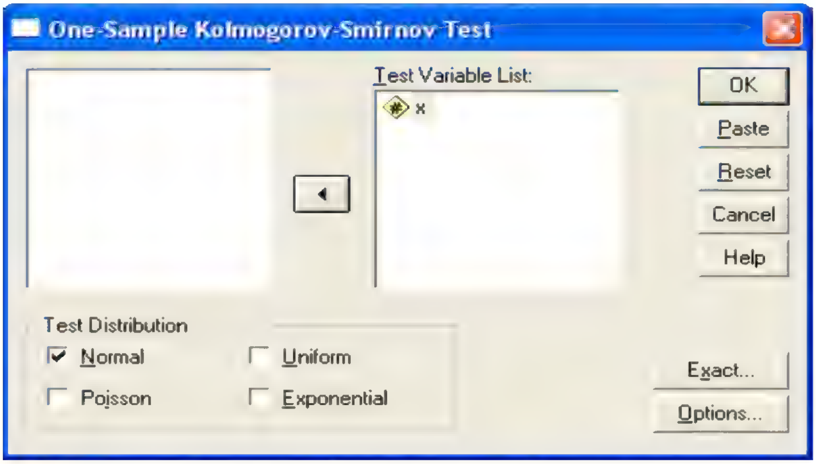
Θέλουμε και πάλι να ελέγξουμε κατά πόσο τα δεδομένα μας προσεγγίζονται από την κανονική κατανομή. Πραγματοποιούμε δηλαδή τον έλεγχο:

Έχουμε ήδη καταχωρίσει τα δεδομένα μας στη μεταβλητή ***x***. Από το menu επιλέγουμε:



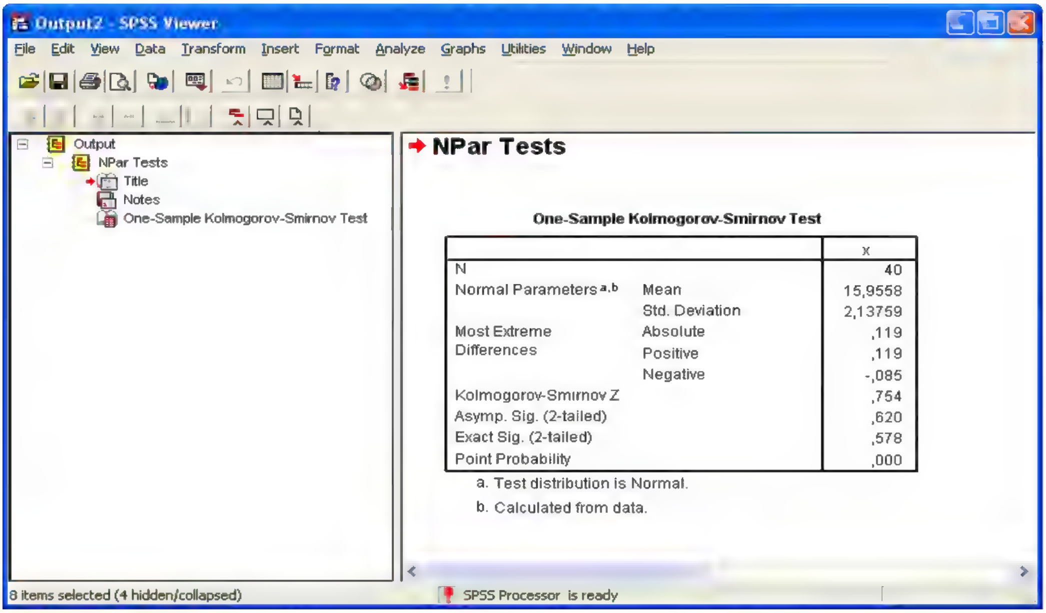
Εικόνα 1.10.1: Επιλογή ελέγχου Kolmogorov–Smirnov ενός δείγματος

Στην καρτέλα **Test Distribution** επιλέγουμε την κατανομή από την οποία θέλουμε να ελέγξουμε ότι προέρχονται τα δεδομένα μας. Στη συγκεκριμένη περίπτωση επιλέγουμε **Normal**. Στην καρτέλα **Exact** επιλέγουμε και πάλι την επιλογή **Exact**. Ο λόγος που κάνουμε αυτή την επιλογή είναι ότι σε περίπτωση μικρών δειγμάτων το p-value που μας δίνει το SPSS ενδέχεται να μας οδηγήσει σε λανθασμένα συμπεράσματα. Έτσι χρησιμοποιώντας το **Exact** παίρνουμε αξιόπιστα αποτελέσματα ανεξάρτητα από την κατανομή και το μέγεθος του δείγματος μας. Έχουμε λοιπόν:



Εικόνα 1.10.2: Κεντρικό menu Kolmogorov – Smirnov

Τα αποτελέσματα που παίρνουμε φαίνονται παρακάτω:



One-Sample Kolmogorov-Smirnov Test

		x
N		40
Normal Parameters[a,b]	Mean	15,9558
	Std. Deviation	2,13759
Most Extreme Differences	Absolute	,119
	Positive	,119
	Negative	-,085
Kolmogorov-Smirnov Z		,754
Asymp. Sig. (2-tailed)		,620
Exact Sig. (2-tailed)		,578
Point Probability		,000

a. Test distribution is Normal.
b. Calculated from data.

Εικόνα 1.10.3: Αποτελέσματα ελέγχου Kolmogorov–Smirnov

Παρατηρούμε ότι ο έλεγχος έγινε για την κανονική κατανομή με μέση τιμή μ = 15.9558 και σ^2 = 2.13759. Επειδή το p-value έχει τιμή μεγαλύτερη από 0.05 (p-value = 0.620) δεν απορρίπτουμε τη μηδενική υπόθεση, οπότε μπορούμε να θεωρήσουμε ότι τα δεδομένα μας πράγματι προέρχονται από κανονική κατανομή με τις συγκεκριμένες παραμέτρους.

Τι γίνεται όμως όταν θέλουμε να προκαθορίσουμε εμείς τις παραμέτρους μιας κατανομής; Ας δούμε το παρακάτω παράδειγμα.

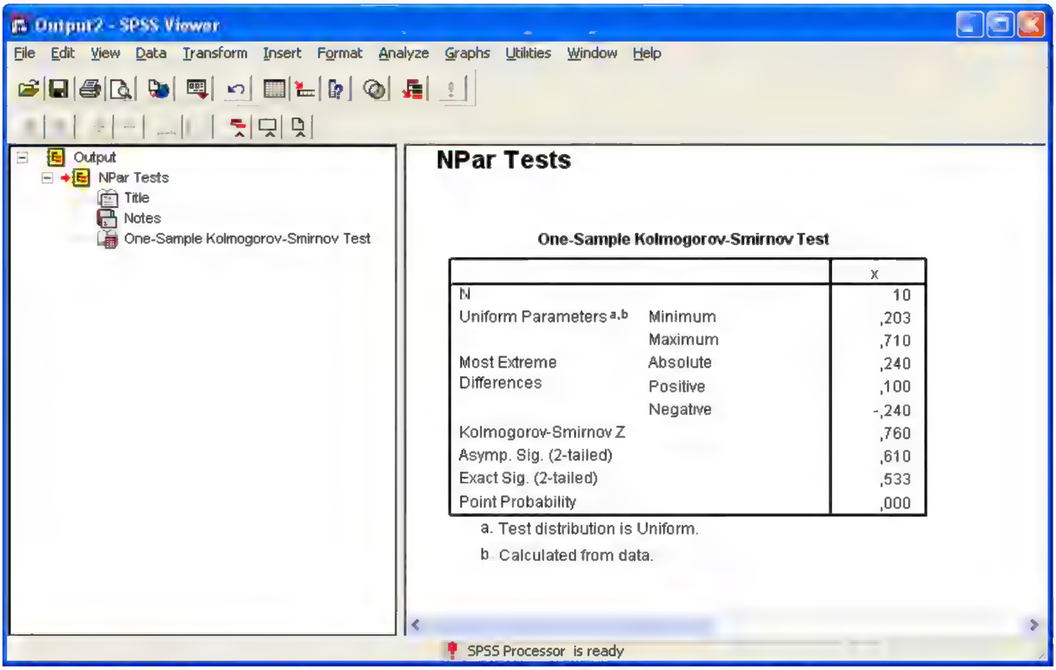
Παράδειγμα

Έχουμε μια συνάρτηση κατανομής F(x)για την οποία παίρνουμε το τυχαίο δείγμα:

0,621	0,503	0,203	0,477	0,710
0,581	0,329	0,480	0,554	0,382

Θέλουμε να ελέγξουμε την υπόθεση ότι το δείγμα έχει προέλθει από την ομοιόμορφη κατανομή στο διάστημα (0,1).

Αν ακολουθούσαμε την προηγούμενη διαδικασία θα καταλήγαμε στο εξής:



One-Sample Kolmogorov-Smirnov Test

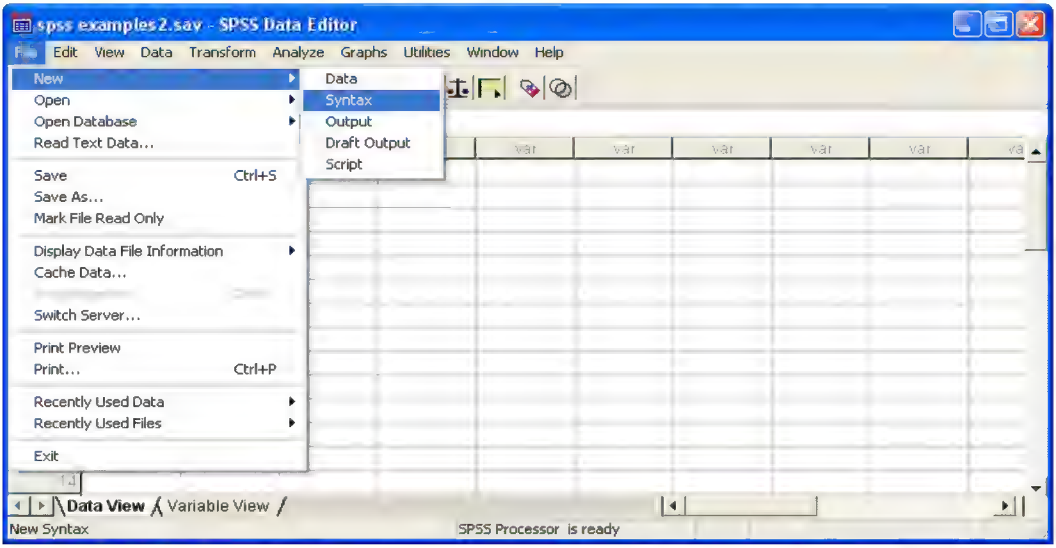
		x
N		10
Uniform Parameters a,b	Minimum	,203
	Maximum	,710
Most Extreme Differences	Absolute	,240
	Positive	,100
	Negative	-,240
Kolmogorov-Smirnov Z		,760
Asymp. Sig. (2-tailed)		,610
Exact Sig. (2-tailed)		,533
Point Probability		,000

a. Test distribution is Uniform.

b. Calculated from data.

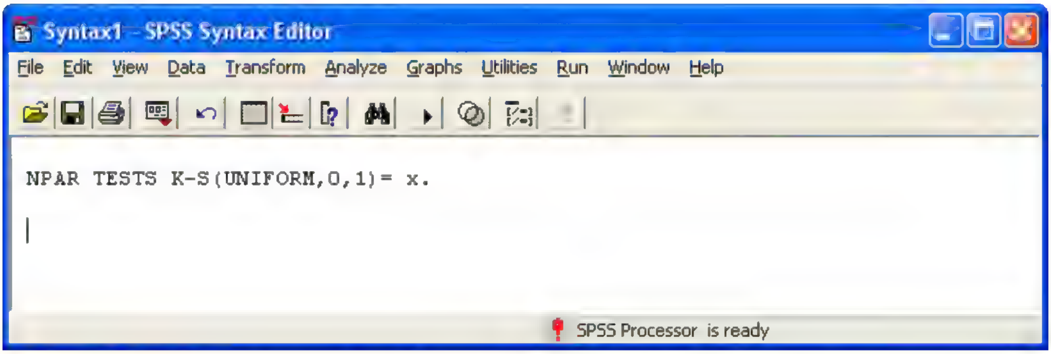
Εικόνα 1.10.4: Αποτελέσματα ελέγχου Kolmogorov–Smirnov

Παρατηρούμε ότι οι παράμετροι που ορίστηκαν για την ομοιόμορφη κατανομή δεν είναι τα (0,1) που θέλαμε αλλά, τα (0.203, 0.710) που προήλθαν από τα δεδομένα μας (ελάχιστη και μέγιστη τιμή). Για να πραγματοποιήσουμε τον έλεγχο που θέλουμε ακολουθούμε τα εξής βήματα:



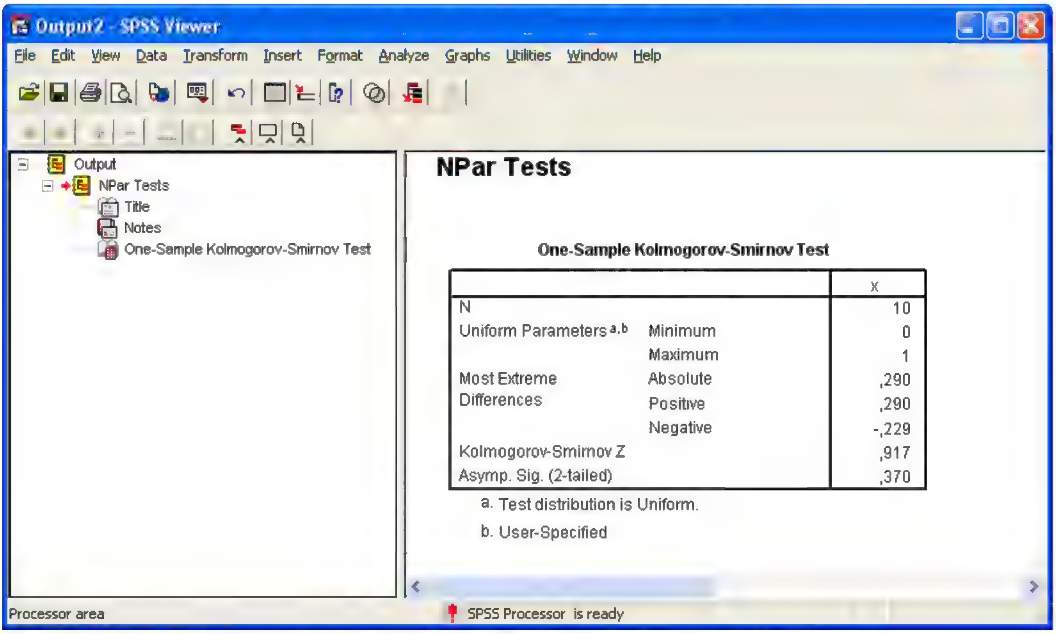
Εικόνα 1.10.5: Άνοιγμα παράθυρου εντολών (SPSS Syntax)

Στο menu εισάγουμε τα εξής:



Εικόνα 1.10.6: Δημιουργία ελέγχου Kolmogorov - Smirnov με συγκεκριμένες παραμέτρους

Από το menu επιλέγω **Run** και παίρνω τα εξής αποτελέσματα:



		x
N		10
Uniform Parameters[a,b]	Minimum	0
	Maximum	1
Most Extreme Differences	Absolute	,290
	Positive	,290
	Negative	-,229
Kolmogorov-Smirnov Z		,917
Asymp. Sig. (2-tailed)		,370

Εικόνα 1.10.7: Αποτελέσματα ελέγχου Kolmogorov – Smirnov

Πράγματι οι παράμετροι για την κατανομή μας είναι (0,1). Το p-value είναι 0.370, τιμή μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση, άρα μπορούμε να θεωρήσουμε ότι το δείγμα μας προσεγγίζεται από την ομοιόμορφη κατανομή με παραμέτρους 0 και 1.

1.11 Συσχέτιση

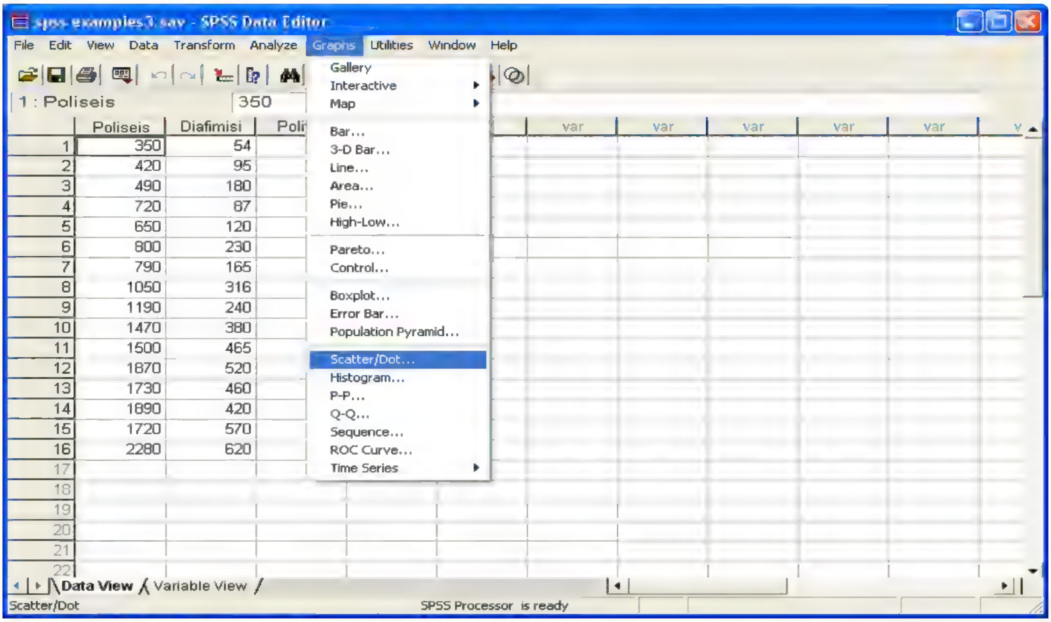
Στον παρακάτω πίνακα παρουσιάζονται κάποια στοιχεία της εταιρίας (Πωλήσεις, Έξοδα Διαφήμισης σε .000€ και Αριθμός Πωλητών) από το έτος ίδρυσης της μέχρι σήμερα.

Έτος	Πωλήσεις (σε .000€)	Έξοδα Διαφήμισης (σε .000€)	Αριθμός Πωλητών
1993	350	54	32
1994	420	95	47
1995	490	180	23
1996	720	87	68
1997	650	120	32
1998	800	230	17
1999	790	165	58
2000	1050	316	75
2001	1190	240	98
2002	1470	380	43
2003	1500	465	76
2004	1870	520	89
2005	1730	460	108
2006	1890	420	76
2007	1720	570	65
2008	2280	620	93

Πίνακας 1.11.1: Πίνακας Δεδομένων

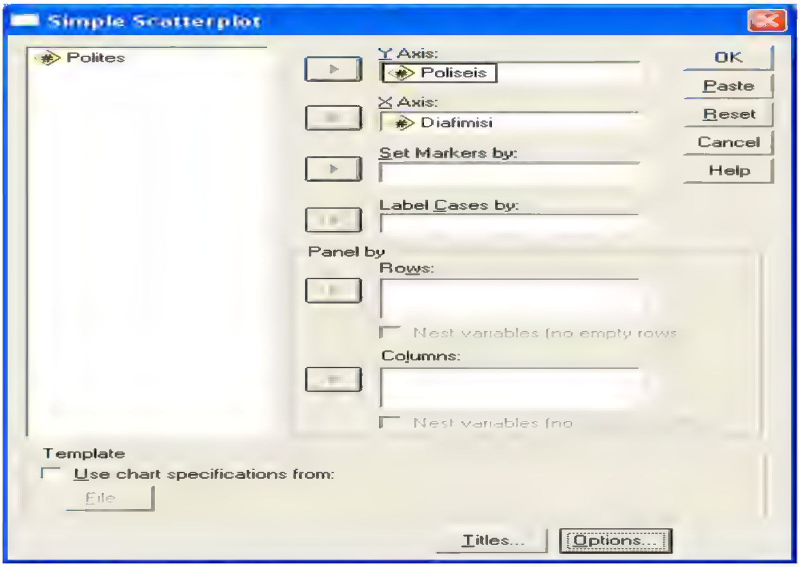
Θα κατασκευάσουμε δύο γραφήματα. Ένα για κάθε εξαρτημένη μεταβλητή. Οι πωλήσεις είναι η ανεξάρτητη μεταβλητή (X) και θα τοποθετηθεί στον άξονα Y. Τα έξοδα διαφήμισης (Y1) και ο αριθμός των πωλητών (Y2) θα είναι οι εξαρτημένες μας μεταβλητές.

Από το menu επιλέγω την καρτέλα **Simple Scatter**.



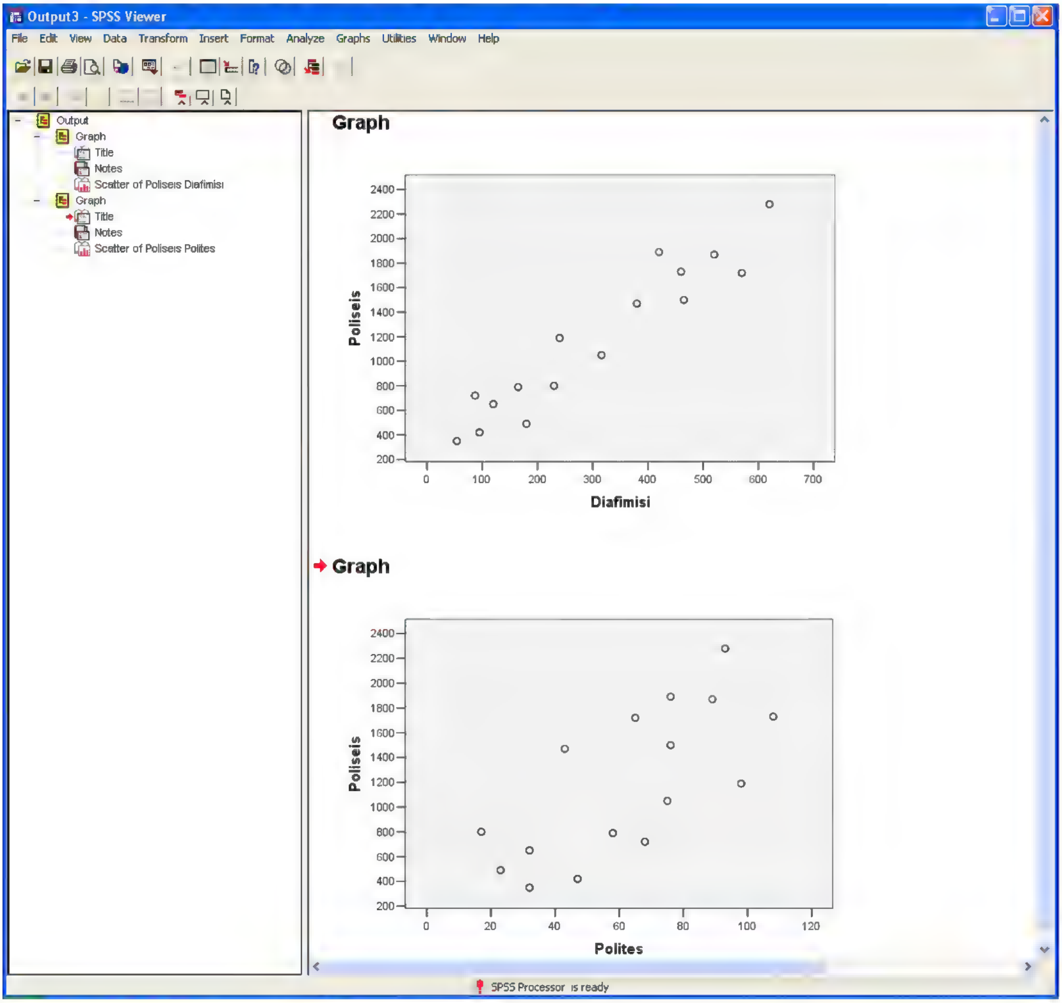
Εικόνα 1.11.1: Επιλογή Scatter Plot

Στον Υ άξονα εισάγω τη μεταβλητή **Poliseis** και στον Χ άξονα πρώτα τη μία μεταβλητή (έστω την **Diafimisi**) και μετά την άλλη (**Polites**)



Εικόνα 1.11.2: Καθορισμός εξαρτημένης και ανεξάρτητης μεταβλητής

Τα γραφήματα που παίρνουμε είναι της μορφής:



Εικόνα 1.11.3: Διαγράμματα Διασποράς

Όπως παρατηρούμε και από τα δύο γραφήματα φαίνεται να υπάρχει κάποια σχέση μεταξύ εξαρτημένων και ανεξάρτητης μεταβλητής. Και στις δύο περιπτώσεις φαίνεται ότι οι πωλήσεις αυξάνονται όσο αυξάνονται τα έξοδα διαφήμισης, αλλά και ο αριθμός των πωλητών. Η διαφορά είναι ότι φαίνεται η σχέση ανάμεσα στις πωλήσεις–έξοδα διαφήμισης να είναι πιο ισχυρή από ότι εκείνη ανάμεσα στις πωλήσεις–αριθμός πωλητών.

Υπάρχουν μέτρα που μας επιτρέπουν να εκτιμήσουμε με μεγάλη ακρίβεια τη συσχέτιση μεταξύ δύο μεταβλητών, αλλά και την ισχύ της συσχέτισης, καθώς επίσης και το είδος αυτής.

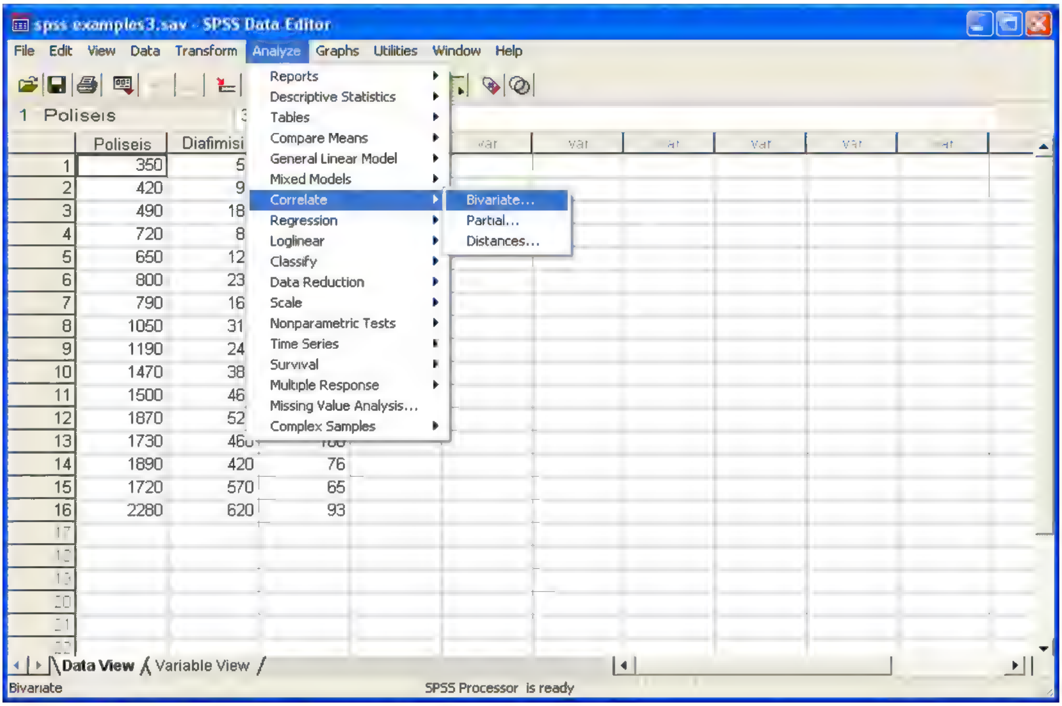
Παράδειγμα

Στο παράδειγμα με την εταιρία που καλείται να επιλέξει την πολιτική που θα ακολουθήσει έχουμε τα εξής δεδομένα:

Έτος	Πωλήσεις (σε .000€)	Έξοδα Διαφήμισης (σε .000€)	Αριθμός Πωλητών
1993	350	54	32
1994	420	95	47
1995	490	180	23
1996	720	87	68
1997	650	120	32
1998	800	230	17
1999	790	165	58
2000	1050	316	75
2001	1190	240	98
2002	1470	380	43
2003	1500	465	76
2004	1870	520	89
2005	1730	460	108
2006	1890	420	76
2007	1720	570	65
2008	2280	620	93

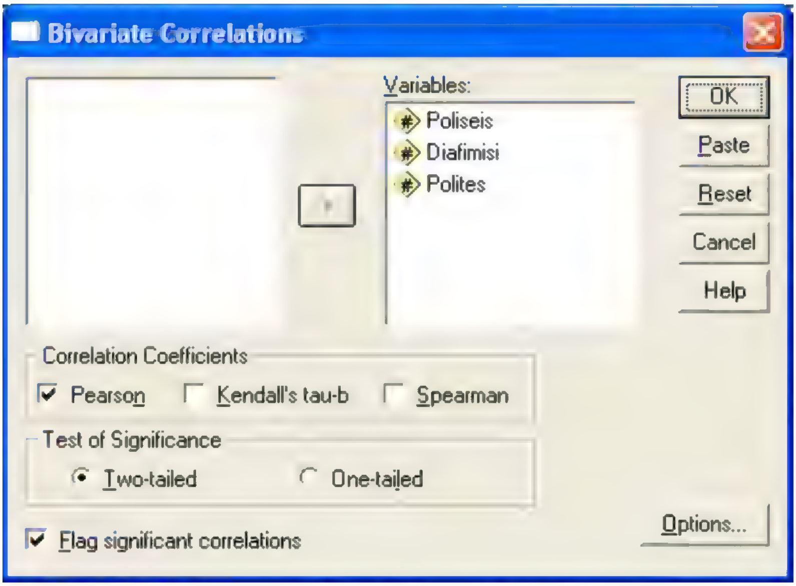
Πίνακας 1.11.1: **Πίνακας Δεδομένων**

Αφού με τα γραφήματα δεν είμαστε σίγουροι ποια είναι η καλύτερη επιλογή θα χρησιμοποιήσουμε το στατιστικό μέτρο r. Από το menu επιλέγουμε:



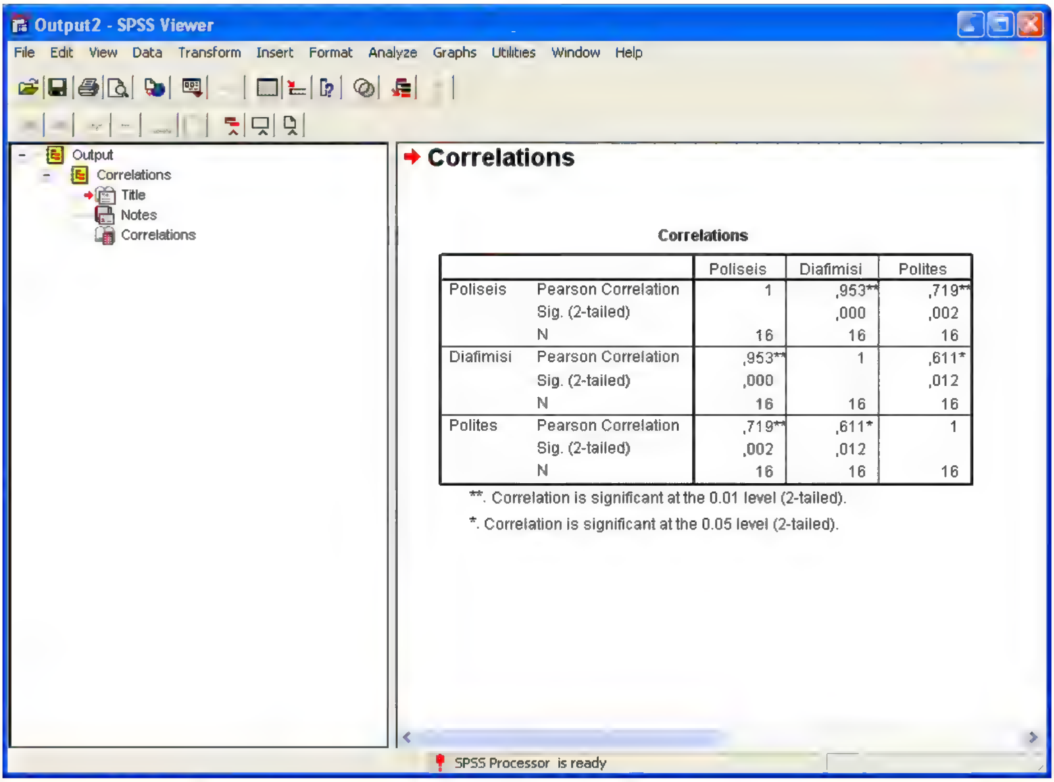
Εικόνα 1.11.4: Επιλογή συντελεστή συσχέτισης

Στη συνέχεια επιλέγουμε τις μεταβλητές για τις οποίες θα βρούμε το συντελεστή συσχέτισης.



Εικόνα 1.11.5: Κεντρικό menu συντελεστή συσχέτισης

Το αποτέλεσμα που παίρνουμε είναι:



Correlations

		Poliseis	Diafimisi	Polites
Poliseis	Pearson Correlation	1	,953**	,719**
	Sig. (2-tailed)		,000	,002
	N	16	16	16
Diafimisi	Pearson Correlation	,953**	1	,611*
	Sig. (2-tailed)	,000		,012
	N	16	16	16
Polites	Pearson Correlation	,719**	,611*	1
	Sig. (2-tailed)	,002	,012	
	N	16	16	16

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Εικόνα 1.11.6: Αποτέλεσμα συντελεστή συσχέτισης

Παρατηρούμε ότι η συσχέτιση μεταξύ της μεταβλητής **Poliseis** και **Diafimisi** είναι πολύ υψηλή (r=0.95), ενώ αντίθετα η συσχέτιση μεταξύ **Poliseis** και **Polites** είναι λιγότερο ισχυρή (r=0.719). Επομένως την εταιρία τη συμφέρει να προχωρήσει σε αύξηση των εξόδων διαφήμισης προκειμένου να αυξήσει περαιτέρω την κερδοφορία της.

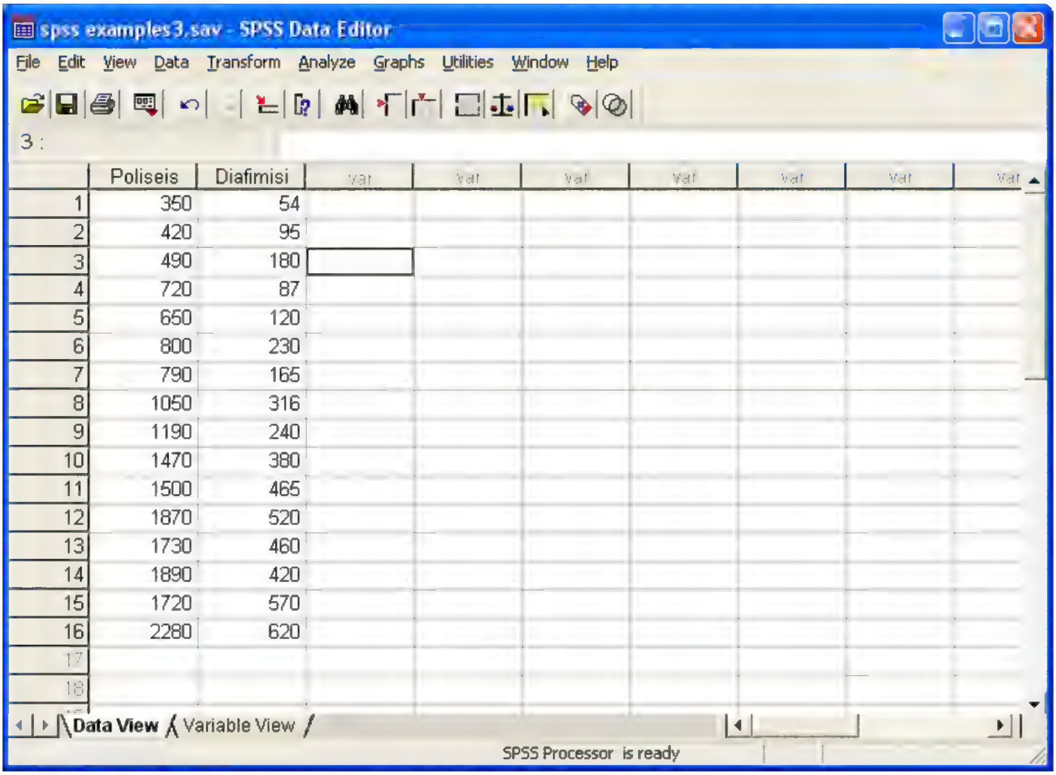
Παρατηρούμε επίσης, ότι τα p-values (**Sig. (2-tailed)**) για τον κάθε συντελεστή είναι μικρότερα από 0.05. Επομένως ισχύει ότι οι συντελεστές συσχέτισης είναι στατιστικά σημαντικοί και άρα απορρίπτουμε τη μηδενική υπόθεση ότι ρ=0.

1.12 Απλή και Πολλαπλή Γραμμική Παλινδρόμηση με χρήση SPSS

Σε αυτήν την ενότητα θα δούμε πώς χειριζόμαστε το SPSS για να επιλύσουμε ένα πρόβλημα απλής γραμμικής παλινδρόμησης. Θα χρησιμοποιήσουμε το ίδιο παράδειγμα με την εταιρία που θέλει να αυξήσει την κερδοφορία της. Είδαμε ότι το καλύτερο μοντέλο παλινδρόμησης προέκυψε από την επιλογή της «Έξοδα Διαφήμισης» ως την ανεξάρτητη μας μεταβλητή, οπότε με αυτήν θα ασχοληθούμε εδώ. Δηλαδή, το μοντέλο μας θα είναι της μορφής:

«Μέσες Πωλήσεις» = a+b«Έξοδα Διαφήμισης»

Καταχωρούμε τα δεδομένα σε δύο στήλες όπως φαίνεται παρακάτω.

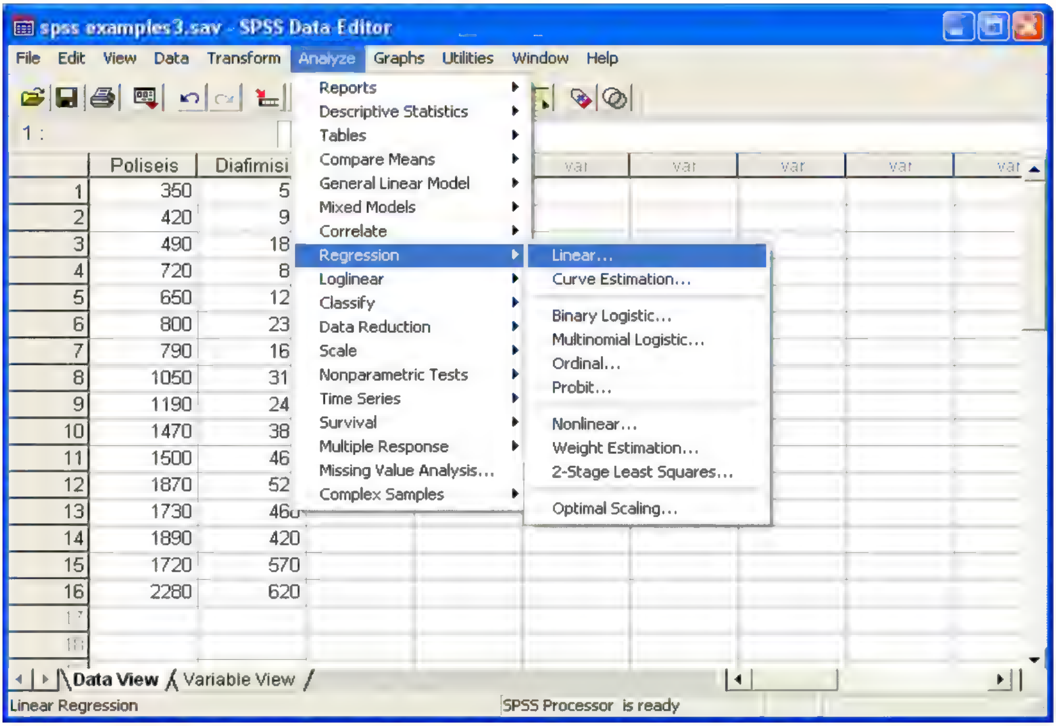


	Poliseis	Diafimisi
1	350	54
2	420	95
3	490	180
4	720	87
5	650	120
6	800	230
7	790	165
8	1050	316
9	1190	240
10	1470	380
11	1500	465
12	1870	520
13	1730	460
14	1890	420
15	1720	570
16	2280	620

Εικόνα 1.12.1: Καταχώρηση δεδομένων παραδείγματος

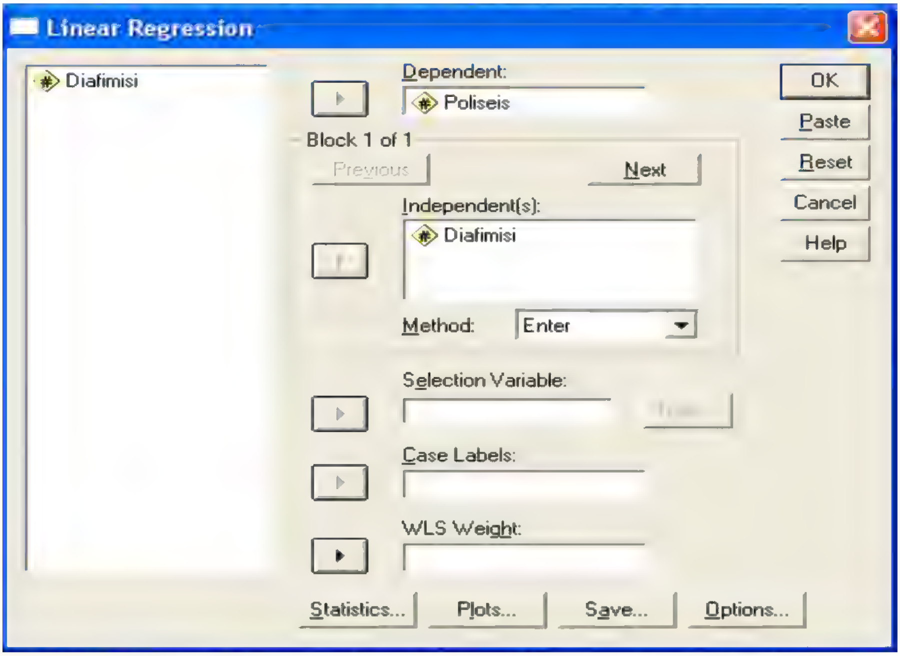
Η μεταβλητή **Poliseis** αφορά την εξαρτημένη μας μεταβλητή και η **Diafimisi** την ανεξάρτητη. Το επόμενο βήμα είναι να δούμε αν μπορούμε να προχωρήσουμε στην κατασκευή του μοντέλου παλινδρόμησης, αν δηλαδή, ικανοποιούνται οι προϋποθέσεις εφαρμογής του.

Αναλυτικά, ακολουθούμε τα βήματα:



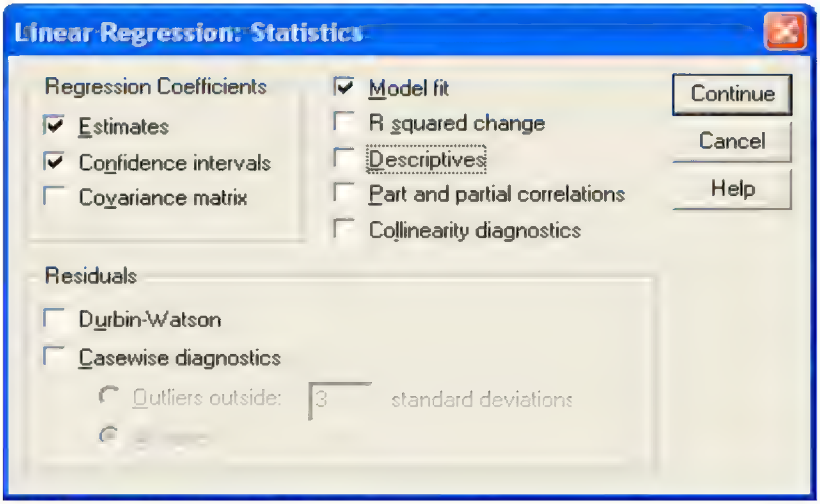
Εικόνα 1.12.2: Επιλογή μοντέλου Απλής Γραμμικής Παλινδρόμησης

Επιλέγουμε την εξαρτημένη (**Poliseis**) και την ανεξάρτητη μεταβλητή (**Diafimisi**):



Εικόνα 1.12.3: Επιλογή εξαρτημένης και ανεξάρτητης μεταβλητής

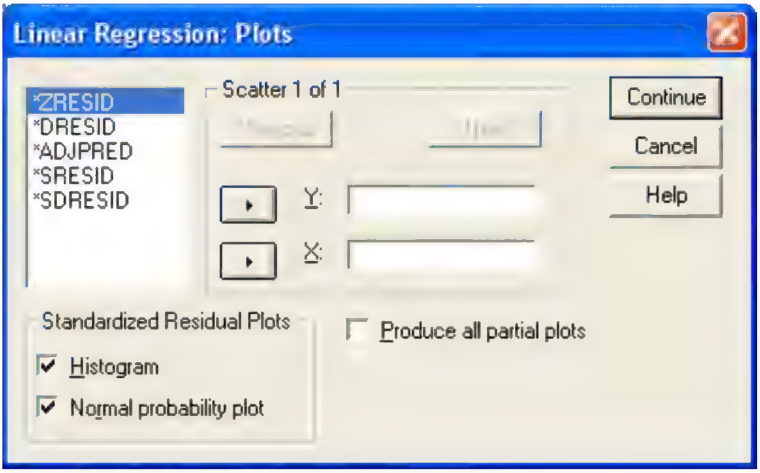
Στη συνέχεια από την καρτέλα **Statistics** επιλέγουμε τα εξής:



Εικόνα 1.12.4: Καρτέλα Statistics

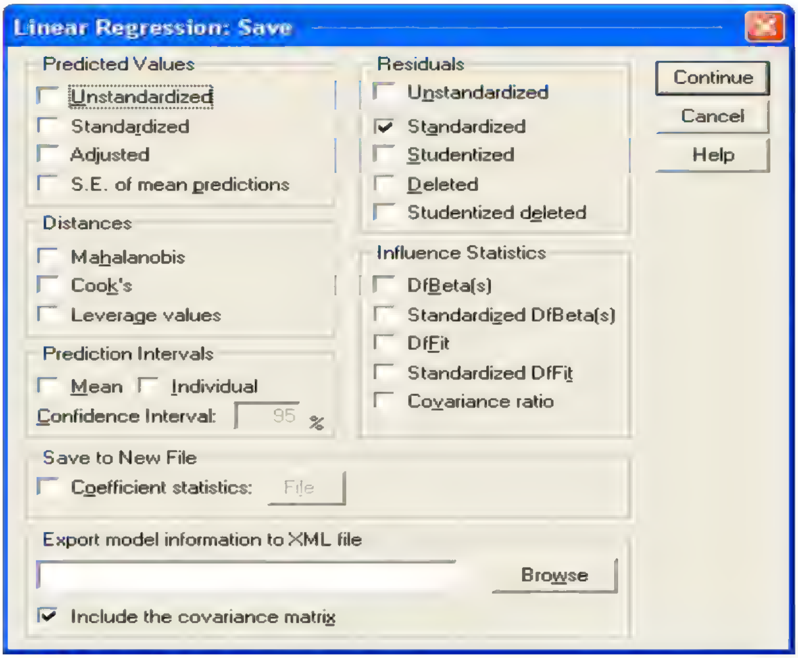
Αν θέλουμε να δούμε για κάθε τιμή του Χ και Υ τα κατάλοιπα αλλά και τις εκτιμηθείσες τιμές του Υ επιλέγουμε και το **Casewise diagnostics**.

Στη συνέχεια από την καρτέλα Plots επιλέγουμε τα διαγράμματα των καταλοίπων όπως φαίνεται παρακάτω:



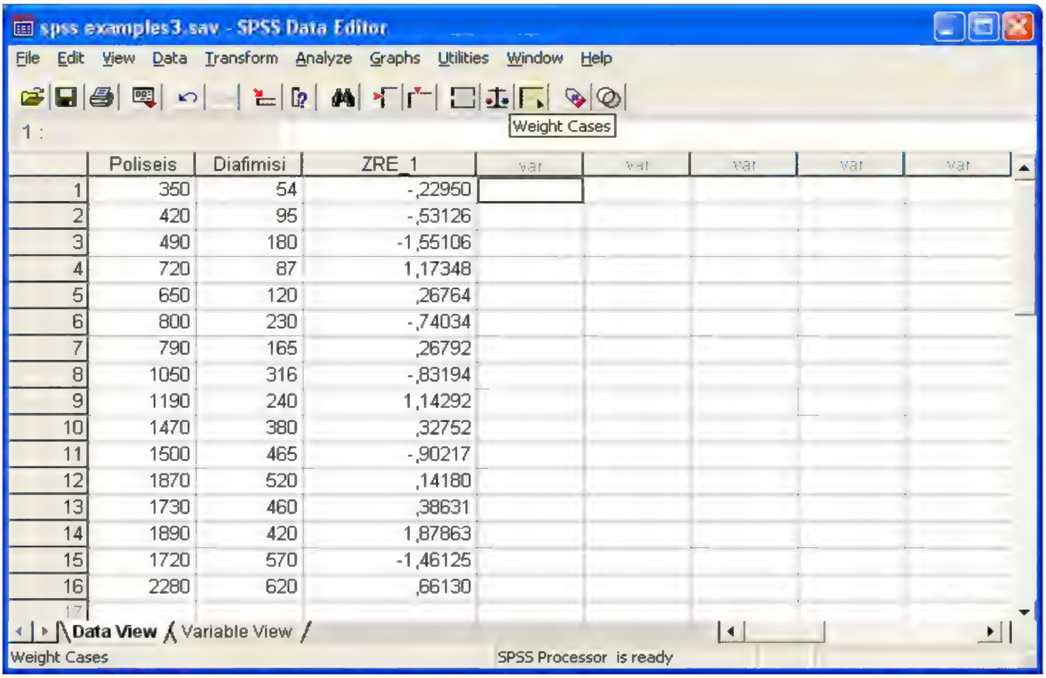
Εικόνα 1.12.5: Καρτέλα Plots

Στην καρτέλα **Save** καθορίζουμε τι θέλουμε να αποθηκευθεί στο κεντρικό menu των μεταβλητών μας. Επιλέγουμε τα **standardized residuals**.



Εικόνα 1.12.6: Καρτέλα Save

Πλέον, στο κεντρικό menu έχει προστεθεί μία ακόμη μεταβλητή που αφορά τα κατάλοιπα:



	Poliseis	Diafimisi	ZRE_1	var	var	var	var	var
1	350	54	-,22950					
2	420	95	-,53126					
3	490	180	-1,55106					
4	720	87	1,17348					
5	650	120	,26764					
6	800	230	-,74034					
7	790	165	,26792					
8	1050	316	-,83194					
9	1190	240	1,14292					
10	1470	380	,32752					
11	1500	465	-,90217					
12	1870	520	,14180					
13	1730	460	,38631					
14	1890	420	1,87863					
15	1720	570	-1,46125					
16	2280	620	,66130					
17								

Εικόνα 1.12.7: Κεντρικό menu μεταβλητών

Όπως είπαμε θα χρησιμοποιήσουμε τα κατάλοιπα για να δούμε αν μπορούμε να προχωρήσουμε με το μοντέλο της παλινδρόμησης. Γενικά για την εφαρμογή ενός μοντέλου παλινδρόμησης πρέπει να ισχύουν τα εξής για τα κατάλοιπα:

√ Ανεξαρτησία
√ Ομοσκεδαστικότητα
√ Κανονικότητα

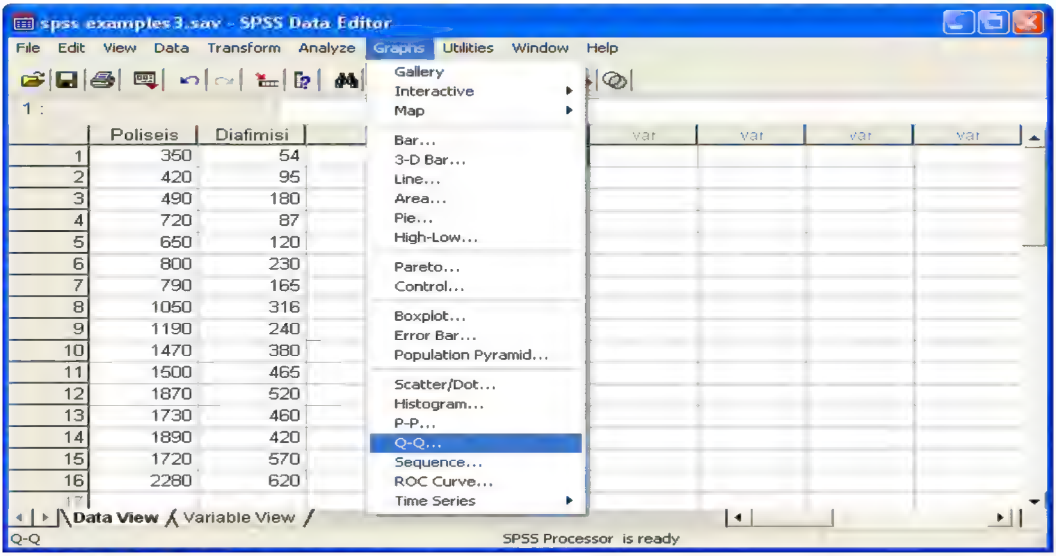
Για τον έλεγχο της ανεξαρτησίας θα χρησιμοποιήσουμε το Runs Test. Το αποτέλεσμα που παίρνουμε:

Runs Test

	Standardized Residual
Test Value [a]	,20472
Cases < Test Value	8
Cases >= Test Value	8
Total Cases	16
Number of Runs	10
Z	,259
Asymp. Sig. (2-tailed)	,796
Exact Sig. (2-tailed)	,810
Point Probability	,190

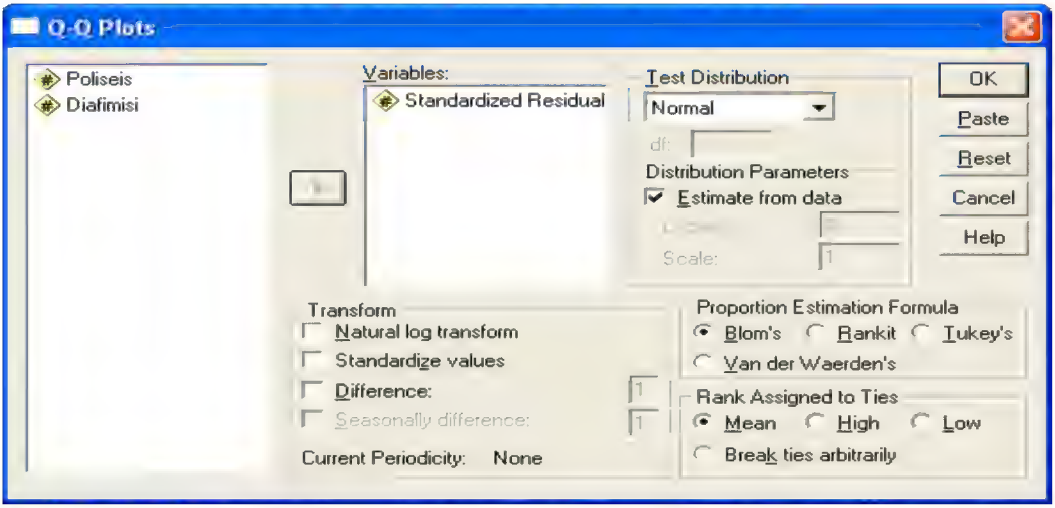
Εικόνα 1.12.8: Runs Test για έλεγχο ανεξαρτησίας καταλοίπων

Το p-value έχει τιμή 0.810, μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση περί ανεξαρτησίας-τυχαιότητας των καταλοίπων. Για τις υπόλοιπες υποθέσεις έχουμε:

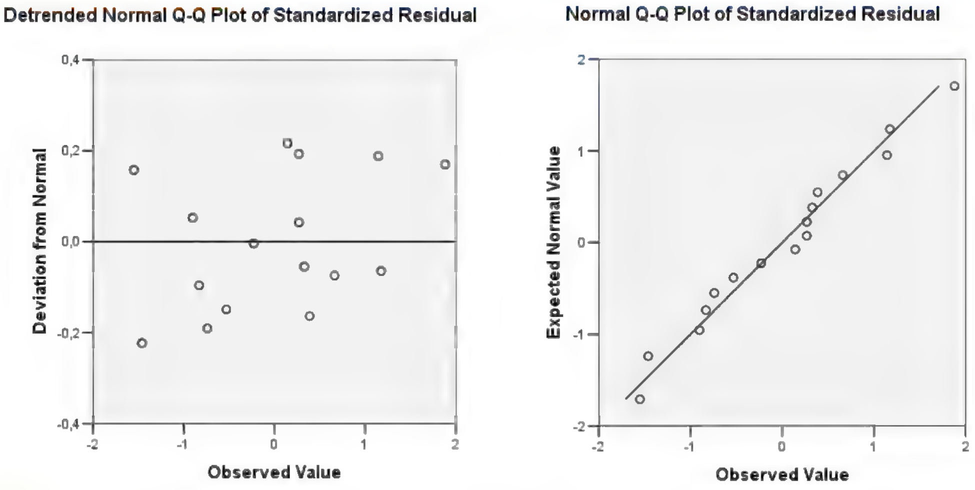


Εικόνα 1.12.9: Επιλογή ελέγχου καταλοίπων

Επιλέγουμε τη μεταβλητή μας και παίρνουμε το εξής γράφημα:



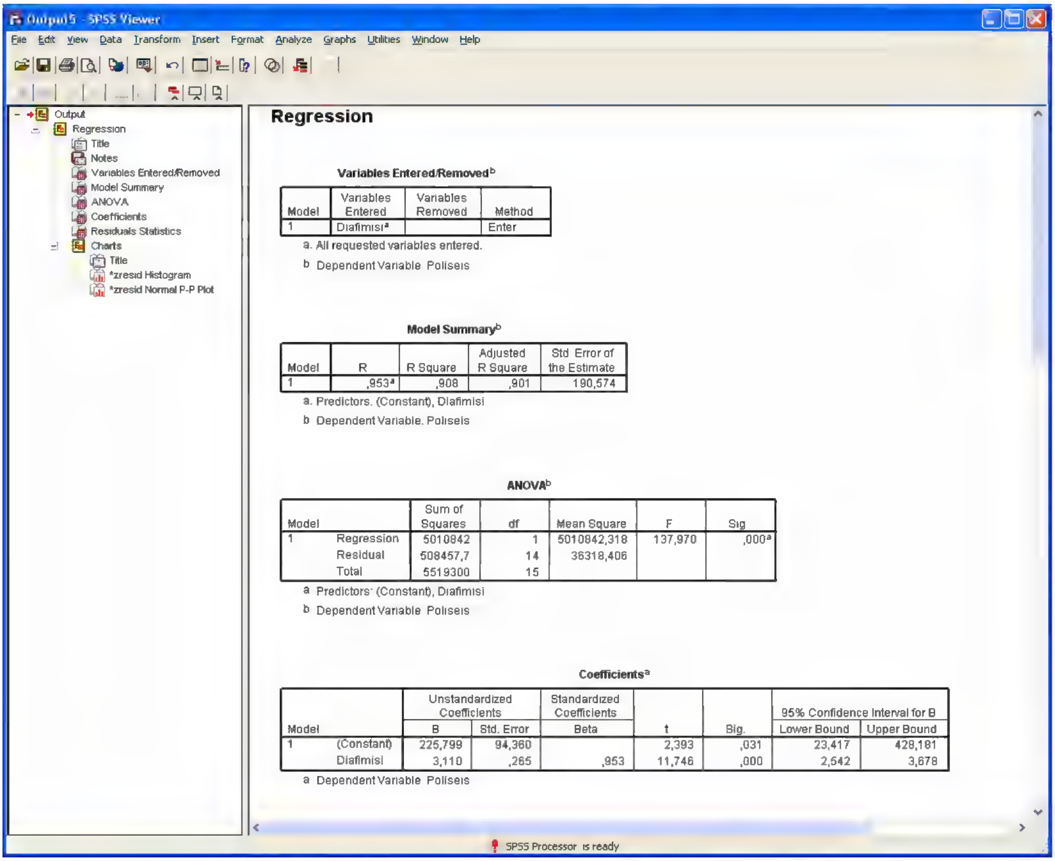
Εικόνα 1.12.10: Επιλογή καταλοίπων για κατασκευή γραφημάτων



Εικόνα 1.12.11: Διαγράμματα καταλοίπων για έλεγχο κανονικότητας και ομοσκεδαστικότητας

Από τα πρώτο γράφημα φαίνεται ότι τα κατάλοιπα κατανέμονται τυχαία γύρω από το μηδέν και επομένως φαίνεται να ισχύει η ομοσκεδαστικότητα. Επίσης, στο δεύτερο γράφημα φαίνεται ότι τα κατάλοιπα δεν απέχουν πολύ από τη γραμμή που δείχνει την κανονικότητα και επομένως μπορούμε να πούμε ότι ισχύει και αυτή η υπόθεση.

Αφού λοιπόν ισχύουν οι παραπάνω υποθέσεις μπορούμε να προχωρήσουμε σε ανάλυση του μοντέλου που προέκυψε.



Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Diafimisi[a]		Enter

a. All requested variables entered.
b. Dependent Variable: Poliseis

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,953[a]	,908	,901	190,574

a. Predictors: (Constant), Diafimisi
b. Dependent Variable: Poliseis

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5010842	1	5010842,318	137,970	,000[a]
	Residual	508457,7	14	36318,406		
	Total	5519300	15			

a. Predictors: (Constant), Diafimisi
b. Dependent Variable: Poliseis

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	225,799	94,360		2,393	,031	23,417	428,181
	Diafimisi	3,110	,265	,953	11,746	,000	2,542	3,678

a. Dependent Variable: Poliseis

Εικόνα 1.12.12: Μοντέλο Απλής Γραμμικής Παλινδρόμησης

Αναλυτικά:

Model Summury

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,953[a]	,908	,901	190,574

Εικόνα 1.12.13: Έλεγχος γραμμικότητας μοντέλου

Βλέπουμε ότι ο συντελεστής συσχέτισης είναι 0.953, πολύ υψηλός. Επίσης, τόσο το r2 (συντελεστής προσδιορισμού) όσο και το r2adj έχουν τιμή που τείνει στο 1, επομένως έχουμε ενδείξεις για τη γραμμικότητα του μοντέλου. Μάλιστα από τον πίνακα της ANOVA βλέπουμε τον έλεγχο που αφορά αν ο συντελεστής προσδιορισμού είναι στατιστικά σημαντικός. Το p-value είναι πρακτικά ίσο με μηδέν και επομένως απορρίπτουμε τη μηδενική υπόθεση ότι το $r^2 = 0$.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	225,799	94,360		2,393	,031	23,417	428,181
	Diafimisi	3,110	,265	,953	11,746	,000	2,542	3,678

Εικόνα 1.12.14: Τιμές στις παραμέτρους του μοντέλου

Παρατηρούμε ότι το a έχει τιμή 225,799 ενώ το b 3,110. Άρα το μοντέλο παλινδρόμησης είναι το:

«Μέσες Πωλήσεις» = 225.799+3,11«Έξοδα Διαφήμισης»

Επίσης, πραγματοποιείται και ο έλεγχος σημαντικότητας για την κάθε μία παράμετρο του μοντέλου. Και για τις δύο παραμέτρους τα p-values είναι μικρότερα από 0.05 (0.031 και ≈0 αντίστοιχα) και επομένως απορρίπτουμε τις μηδενικές υποθέσεις ότι οι τιμές των παραμέτρων a, b είναι μηδέν.

Τέλος, δίνονται και διαστήματα εμπιστοσύνης για την κάθε μία παράμετρο. Δηλαδή, ένα 95% διάστημα εμπιστοσύνης για το a είναι (23.417 , 428.181) ενώ για το b το αντίστοιχο διάστημα εμπιστοσύνης είναι το (2.542 , 3.678).

Πολλαπλή Γραμμική Παλινδρόμηση

Παράδειγμα

Το Υπουργείο Υγείας μιας χώρας θέλει μέσω μιας επιδημιολογικής μελέτης να εξακριβώσει αν μπορεί να προληφθεί ο δείκτης θνησιμότητας σε διάφορες περιοχές της χώρας. Για το λόγο αυτό διεξάγονται πειράματα που έχουν ως αποτέλεσμα την καταγραφή των τιμών συγκεκριμένων μεταβλητών. Παρακάτω φαίνονται τα αποτελέσματα από ένα τυχαίο δείγμα 15 περιοχών από όλη τη χώρα. Ο πίνακας των μεταβλητών παρατίθεται παρακάτω.

Y:	**Δείκτης θνησιμότητας**
X_1:	**Μέση Ετήσια Βροχόπτωση**
X_2:	**Μέση θερμοκρασία Ιανουαρίου**
X_3:	**Μέση θερμοκρασία Ιουλίου**
X_4:	**Διάμεσος του αριθμού των πλήρων χρόνων εκπαίδευσης**
X_5:	**Ποσοστό πληθυσμού που δεν είναι λευκοί**
X_6:	**Σχετικές τιμές διοξειδίου του θείου**

Πίνακας 1.12.1: Πίνακας Μεταβλητών

Ο πίνακας δεδομένων είναι:

Α/Α	Υ	Χ1	Χ2	Χ3	Χ4	Χ5	Χ6
1	942	32,2	21,2	70,9	11,1	1,9	94,6
2	852	17,6	37	69,3	12,3	1,1	13,1
3	924	37,6	36,9	79,3	10,1	10	9,8
4	928	31,7	36,4	72,3	11	2	28,4
5	949	33	34,5	73,1	10,2	5,4	243
6	901	36	32,2	73,2	11,2	0,9	1,4
7	938	39,7	36,7	71,3	10,7	1,7	11
8	901	42,8	35,4	78,8	11,3	2,4	18,5
9	942	10	25,5	75,9	10,7	4	56,4
10	1010	29	31,2	72,2	9,8	5,6	87,6
11	885	42,9	32,2	69,8	11,7	1,1	89,7
12	1039	37,1	34,5	75,7	10,5	35,1	78,7
13	988	54,6	41,2	77,2	12,1	15	129
14	981	37,3	34,5	75,5	9	4	74,1
15	944	37,3	38,4	74,4	11,3	3,6	54,7

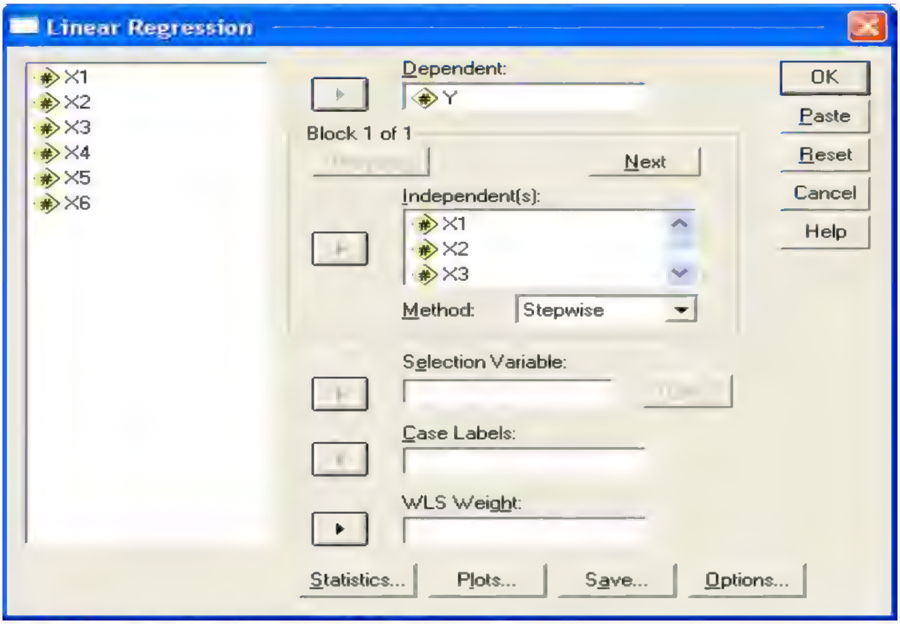
Πίνακας 1.12.2: Πίνακας Δεδομένων πολλαπλής παλινδρόμησης

Το μοντέλο που θα εφαρμόσουμε είναι της μορφής:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Έχοντας τόσες πολλές μεταβλητές προκύπτει το ερώτημα ποιες από όλες αυτές είναι κατάλληλες για να εξηγήσουν την εξαρτημένη μας μεταβλητή. Θα εφαρμόσουμε τη μέθοδο **Stepwise Regression** για να επιλέξουμε το κατάλληλο μοντέλο.

Στο menu της παλινδρόμησης επιλέγουμε τα εξής:



Εικόνα 1.12.15: Επιλογή Stepwise Regression

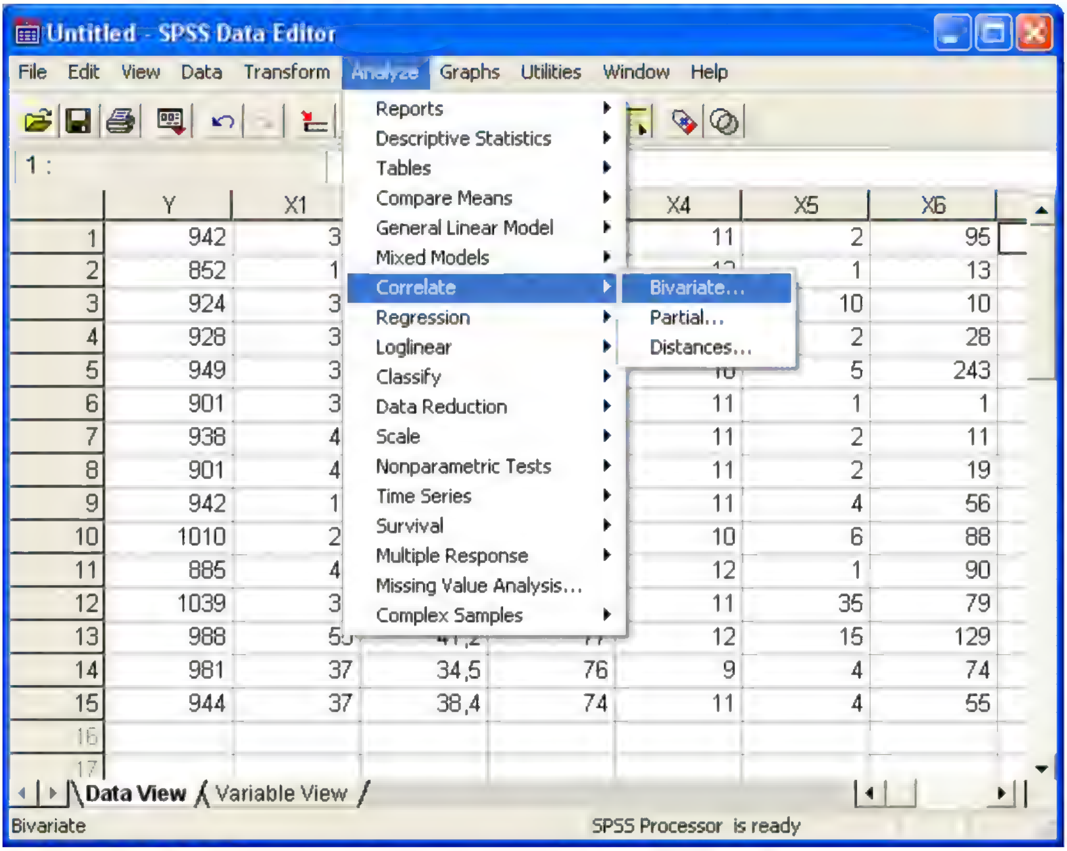
Δηλαδή καθορίσαμε ποιες είναι οι ανεξάρτητες μεταβλητές και ποια η εξαρτημένη και στη συνέχεια στην καρτέλα **Method** επιλέξαμε τη μέθοδο που θέλουμε να ακολουθήσουμε.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	X5	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	X4	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

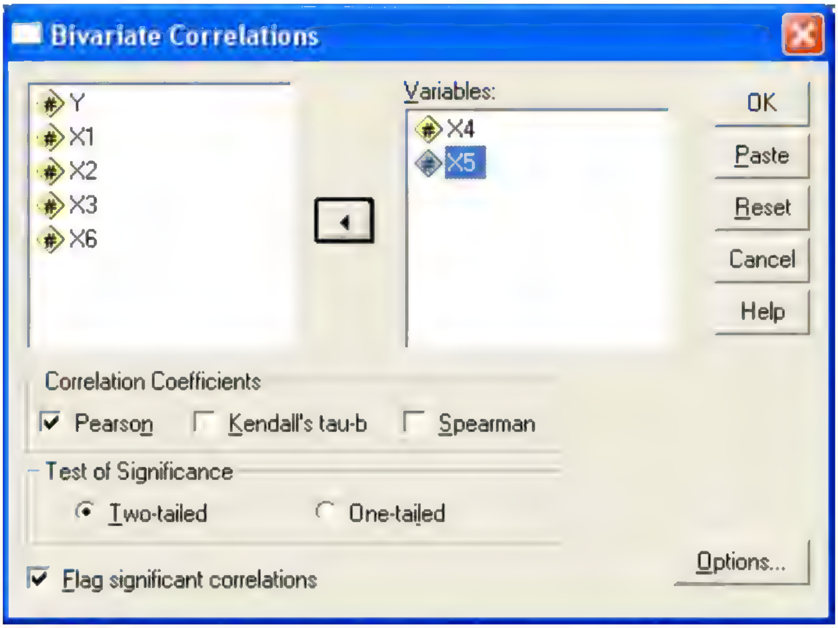
Εικόνα 1.12.16: Αποτέλεσμα Stepwise Regression

Με τη μέθοδο αυτή καταλήξαμε σε μοντέλο με 2 ανεξάρτητες μεταβλητές. Τη μεταβλητή που αφορά τη διάμεσο του αριθμού των πλήρων χρόνων εκπαίδευσης και αυτή που αφορά το ποσοστό των ατόμων που δεν είναι λευκοί. Το επόμενο βήμα είναι να εντοπίσουμε αν έχουμε **πολυσυγγραμμικότητα** στο μοντέλο μας. Επιλέγουμε:



Εικόνα 1.12.17: Έλεγχος πολυσυγγραμμικότητας

Στο κεντρικό menu καθορίζουμε τις μεταβλητές που θέλουμε να ελέγξουμε καθώς επίσης και τους ελέγχους που μας ενδιαφέρουν.



Εικόνα 1.12.18: Κεντρικό menu ελέγχου πολυσυγγραμμικότητας

Το αποτέλεσμα μας δίνει ότι οι δύο μεταβλητές είναι ανεξάρτητες μεταξύ τους αφού το p-value έχει τιμή μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε την υπόθεση περί ανεξαρτησίας των μεταβλητών.

Correlations

		X4	X5
X4	Pearson Correlation	1	-,129
	Sig. (2-tailed)		,647
	N	15	15
X5	Pearson Correlation	-,129	1
	Sig. (2-tailed)	,647	
	N	15	15

Εικόνα 1.12.19: Αποτέλεσμα ελέγχου πολυσυγγραμμικότητας

Το μοντέλο μας λοιπόν έχει τη μορφή:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	917,611	11,518		79,670	,000	892,729	942,494
	X5	3,836	1,088	,699	3,527	,004	1,486	6,186
2	(Constant)	1199,064	97,516		12,296	,000	986,595	1411,533
	X5	3,509	,875	,640	4,009	,002	1,602	5,417
	X4	-25,712	8,869	-,463	-2,899	,013	-45,036	-6,389

Εικόνα 1.12.20: Μοντέλο πολυσυγγραμμικότητας

Η ευθεία παλινδρόμησης είναι η εξής: $\hat{Y} = 1199.06 - 25.7X_4 + 3.51X_5$

Για να δούμε αν μπορούμε να χρησιμοποιήσουμε το μοντέλο μας, πρέπει να ελέγξουμε κάποιες υποθέσεις.

Γραμμικότητα

Ο συντελεστής προσδιορισμού για το μοντέλο μας είναι 0.649 όπως φαίνεται και από τον παρακάτω πίνακα:

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,699[a]	,489	,450	35,999
2	,836[b]	,699	,649	28,734

a. Predictors: (Constant), X5

b. Predictors: (Constant), X5, X4

c. Dependent Variable: Y

Εικόνα 1.12.21: Συντελεστής προσδιορισμού

Θέλουμε να κάνουμε τον έλεγχο:

H_0: Η ευθεία παλινδρόμησης δεν εξηγεί καθόλου τις μεταβλητές
H_1: Η ευθεία παλινδρόμησης εξηγεί ένα μέρος των μεταβλητών

Από τον πίνακα της ANOVA παίρνουμε ότι:

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16120,480	1	16120,480	12,439	,004[a]
	Residual	16847,120	13	1295,932		
	Total	32967,600	14			
2	Regression	23059,961	2	11529,981	13,965	,001[b]
	Residual	9907,639	12	825,637		
	Total	32967,600	14			

a. Predictors: (Constant), X5

b. Predictors: (Constant), X5, X4

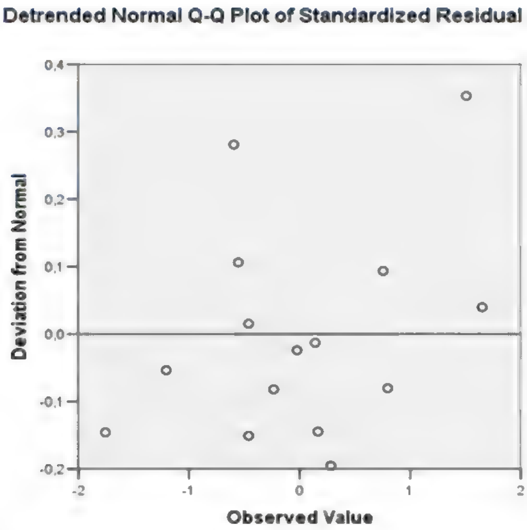
c. Dependent Variable: Y

Εικόνα 1.12.22: Έλεγχος σημαντικότητας Συντελεστής προσδιορισμού

Αφού το p-value έχει τιμή 0.001 < 0.005 απορρίπτουμε τη μηδενική υπόθεση και άρα θεωρούμε ότι υπάρχει γραμμικότητα στο μοντέλο μας.

Ανεξαρτησία και Ομοσκεδαστικότητα

Θα ελέγξουμε την υπόθεση για ανεξαρτησία του μοντέλου μέσου των καταλοίπων. Το γράφημά μας έχει την εξής μορφή:



Εικόνα 1.12.23: Διάγραμμα καταλοίπων για έλεγχο ανεξαρτησίας και ομοσκεδαστικότητας

Φαίνεται να ισχύουν και οι δύο υποθέσεις.

Κανονικότητα

Ο τελευταίος έλεγχος για να μπορέσουμε να προχωρήσουμε με την εφαρμογή του μοντέλου είναι αυτός της κανονικότητας. Θα χρησιμοποιήσουμε τον μη παραμετρικό έλεγχο Komogorov–Smirnov. Παίρνουμε τα εξής αποτελέσματα.

One-Sample Kolmogorov-Smirnov Test

		Standardized Residual
N		15
Normal Parameters[a,b]	Mean	,0000000
	Std. Deviation	,92582010
Most Extreme Differences	Absolute	,127
	Positive	,114
	Negative	-,127
Kolmogorov-Smirnov Z		,492
Asymp. Sig. (2-tailed)		,969
Exact Sig. (2-tailed)		,944
Point Probability		,000

a. Test distribution is Normal.

b. Calculated from data.

Εικόνα 1.12.24: Έλεγχος κανονικότητας Kolmogorv-Smirnov

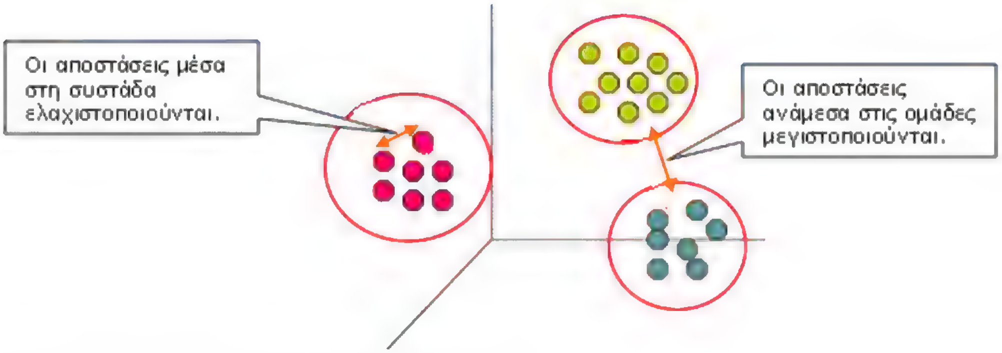
Αφού καταλήξαμε ότι ισχύουν όλες οι απαραίτητες προϋποθέσεις μας, μπορούμε να χρησιμοποιήσουμε το μοντέλο μας για να εξαγάγουμε συμπεράσματα. Βέβαια είναι πολύ σημαντικό να τονίσουμε ότι η πολλαπλή παλινδρόμηση είναι πολύ απαιτητική όσον αφορά τις προϋποθέσεις εφαρμογής της και καλό θα είναι να είμαστε ιδιαίτερα προσεκτικοί στην πλήρη τήρηση όλων των απαιτούμενων προϋποθέσεων ώστε να μπορούμε με ασφάλεια να εξάγουμε συμπεράσματα τα οποία θα είναι και βάσιμα αλλά και θα στηρίζονται σε θεμελιωμένες στατιστικές τεχνικές.

ΠΑΡΑΡΤΗΜΑ ΙΙ: Συσταδοποίηση

1.1 Εισαγωγικά

Το πρόβλημα της συσταδοποίησης σχετίζεται με την τμηματοποίηση (clustering) ενός συνόλου δεδομένων **σε συστάδες** έτσι ώστε τα στοιχεία του συνόλου των δεδομένων που ανήκουν σε μία συστάδα **να είναι περισσότερο όμοια μεταξύ τους από ότι είναι με τα στοιχεία των άλλων συστάδων**. Η συσταδοποίηση είναι μία μέθοδος ευρέως διαδεδομένη σε αρκετούς επιστημονικούς τομείς. Μπορεί να βρεθεί με διαφορετικά ονόματα σε διαφορετικά πεδία, όπως μη εποπτευόμενη μάθηση (unsupervised learning) στην ανάγνωση προτύπων, αριθμητική ταξονομία (numerical taxonomy) στη βιολογία, οικολογία, τυπολογία, στις κοινωνικές επιστήμες και τμηματοποίηση στη θεωρία γραφών και στις Βάσεις Δεδομένων.

Η εύρεση των συστάδων-ομάδων πρέπει να γίνεται με τέτοια τρόπο ώστε να επιτυγχάνεται το ακόλουθο: τα στοιχεία, που εμπεριέχονται σε κάθε ομάδα να είναι όμοια ή να σχετίζονται και διαφορετικά ή να μη συσχετίζονται με τα στοιχεία των άλλων ομάδων, όπως παρουσιάζεται στην παρακάτω εικόνα:



Σχήμα 1.1.1: Απεικόνιση συσταδοποίησης

1.2 Είδη συσταδοποίησης

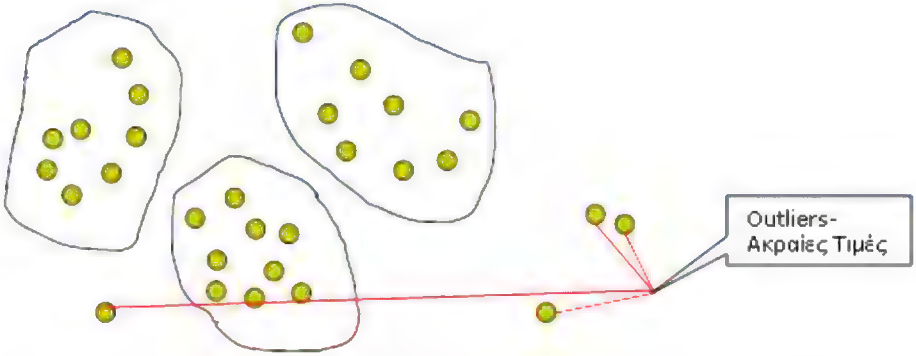
Τα είδη συαταδοποίησης καθώς και ο διαχωρισμός τους, οφείλεται στον τρόπο με τον οποίο πραγματοποιείται η επιλογή των συστάδων. Τα δύο πιο διαδεδομένα είδη συσταδοποίησης είναι **η Διαχωριστική (Partional Clustering) και η Ιεραρχική Συσταδοποίηση (Hierarchical clustering)**.

Στη Διαχωριστική Συσταδοποίηση (Partional Clustering), ο διαμερισμός των αντικειμένων σε μη επικαλυπτόμενες συστάδες γίνεται με τέτοιο τρόπο ώστε κάθε αντικείμενο να ανήκει σε ακριβώς ένα υποσύνολο.

Η επόμενη πιο διαδεδομένη μέθοδος είναι η **Ιεραρχική Συσταδοποίηση** (Hierarchical clustering) στην οποία έχουμε ένα σύνολο από εμφωλευμένες (nested-«φωλιασμένες») συστάδες, όπου επιτρέπουμε σε μία συστάδα να έχει υποσυστάδες οργανωμένες σε ένα ιεραρχικό δέντρο.

Υπάρχουν και άλλα είδη συσταδοποίησης όπως η επικαλυπτόμενη, όπου ένα σημείο ανήκει σε περισσότερες συστάδες. Η ασαφή συσταδοποίηση, στην οποία ένα σημείο-αντικείμενο ανήκει σε κάθε συστάδα με κάποιο βάρος μεταξύ του μηδέν και του ένα, συνήθως τα βάρη για κάθε σημείο αθροίζουν στη μονάδα και η συγκεκριμένη μέθοδος παρουσιάζει πολλά κοινά με τη πιθανοτική συσταδοποίηση. Μία άλλη μέθοδος, που χρησιμοποιείται ευρέος, ειδικά στις περιπτώσεις που έχουμε ενδείξεις ύπαρξης outliers ή μη ενδιαφέρουσας πληροφορίας, είναι η **Μερική-Πλήρης συσταδοποίηση**, όπου ομαδοποιούμε μόνο κάποια από τα δεδομένα.

Ένα από τα σημαντικά προβλήματα της συσταδοποίησης είναι η εύρεση και η αντιμετώπιση **του θορύβου και των outliers**, όταν υπάρχουν ανάμεσα στα στοιχεία-αντικείμενα ακραίες τιμές, που είναι εξαιρέσεις ως προς τις συνηθισμένες ή αναμενόμενες τιμές. (Σχήμα 2)



Σχήμα 1.2.1: Απεικόνιση outliers- ακραίες τιμές στη συσταδοποίηση

1.3 Αλγόριθμοι Συσταδοποίησης

Για να επιτύχουμε την καλύτερη δυνατή δημιουργία συστάδων έχουν δημιουργηθεί αρκετοι αλγόριθμοι συσταδοποίησης. Στο παρόν παράρτημα θα αναλύσουμε τους δύο βασικούς αλγόριθμους, τον αλγόριθμο K-means και την Ιεραρχική συσταδοποίηση.
Ο K- means είναι ένας διαχωριστικός αλγόριθμος, όπου κάθε συστάδα σχετίζεται με ένα κεντρικό σημείο (centroid) και κάθε σημείο ανατίθεται στη συστάδα με το κοντινότερο κεντρικό σημείο. Τα αρχικά κεντρικά σημεία συνήθως επιλέγονται τυχαία και καθώς «τρέχουμε» τον K-means αλγόριθμο, παράγονται οι τελικές συστάδες και τα αρχικά κέντρα διαφοροποιούνται. Όταν εφαρμόζουμε τον K-means αλγόριθμο εκτελούνται τα ακόλουθα βήματα:

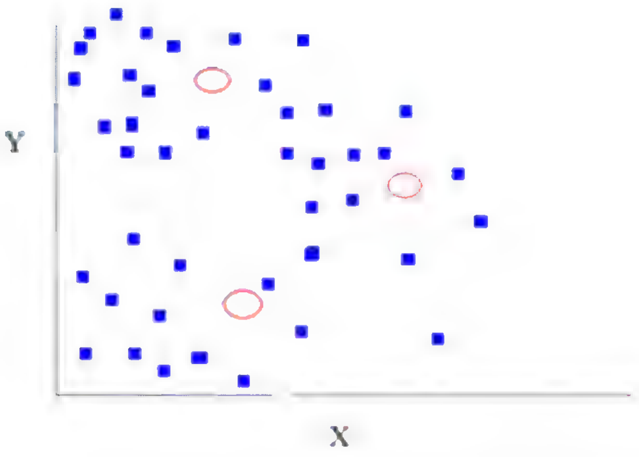
1. Αρχικά επιλέγουμε τα K σημεία, που αποτελούν τα αρχικά κεντρικά σημεία.

2. Τα σημεία που βρίσκονται πιο κοντά στα K σημεία συμπεριλαμβάνονται σε μία συστάδα (συνήθως η απόσταση που χρησιμοποιούμε είναι η Ευκλείδεια απόσταση).

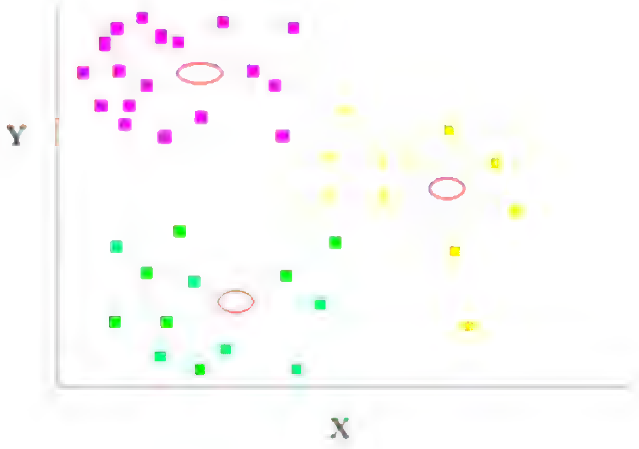
3. Υπολογίζονται εκ νέου τα κεντρικά σημεία (συνήθως τα κεντρικά σημεία είναι το μέσο των σημείων της συστάδας) κάθε συστάδας.

4. Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι τα κεντρικά σημεία να μη μεταβάλλονται.

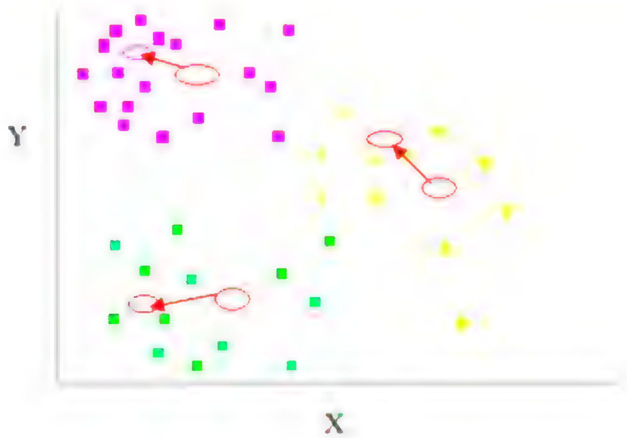
Στα παρακάτω σχήματα απεικονίζονται τα βήματα του αλγόριθμου K-means, για να κατανοήσετε καλύτερα τη λειτουργία του. Με k_1,k_2,k_3 απεικονίζονται τα κεντρικά σημεία και το πλήθος των συστάδων που επιθυμούμε να δημιουργήσουμε είναι τρεις.



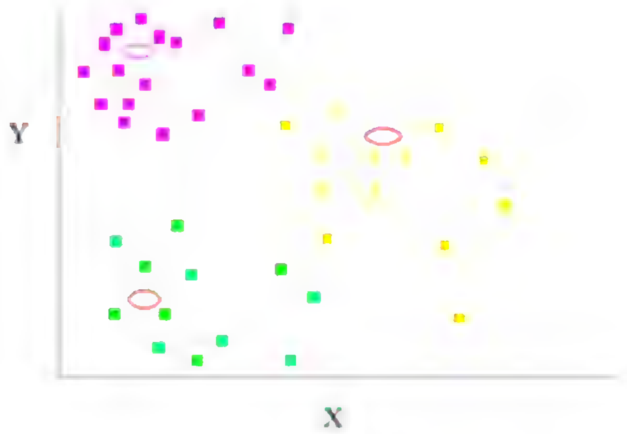
Σχήμα 1.3.1: Αρχική κατάσταση-Ορισμός αρχικών κεντρικών σημείων



Σχήμα 1.3.2: Τα αντικείμενα-στοιχεία ανατίθενται στο πιο γειτονικό από τα τρία αρχικά σημεία

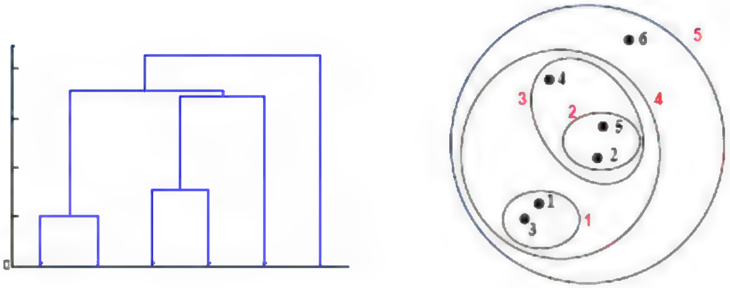


Σχήμα 1.3.3: Επανυπολογισμός του κέντρου βάρους κάθε σημείου



Σχήμα 1.3.4: Τελική απεικόνιση

Στη συνέχεια, θα ασχοληθούμε με την **Ιεραρχική Συσταδοποίηση**. Στην Ιεραρχική Συσταδοποίηση παράγουμε ένα σύνολο από εμφωλευμένες συστάδες σε ένα ιεραρχικό δέντρο, που αποκαλούμε δενδρόγραμμα. Το δενδρόγραμμα είναι ένα διάγραμμα, που μοιάζει με δένδρο και καταγράφει τις ακολουθίες από συγχώνευσεις και διαχωρισμούς ανάμεσα στα αντικείμενα-στοιχεία, που θα αποτελέσουν τη συστάδα. Το βασικό πλεονέκτημα της Ιεραρχικής συσταδοποίησης είναι πως δε χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό συστάδων, αλλά μπορεί να επιτευχθεί ο επιθυμητός αριθμός συστάδων κόβοντας νοητά το δενδρόγραμμα εκεί που επιθυμούμε.



Σχήμα 1.3.5: Σχηματική απεικόνιση Ιεραρχικής Συσταδοποίησης

Υπάρχουν δύο βασικοί τύποι Ιεραρχικής συσταδοποίησης. Ο **Συσσωρευτικός τύπος (Agglomerative)**, όπου αρχίζουμε με όλα τα σημεία ως ξεχωριστές συστάδες και σε κάθε βήμα συγχωνεύουμε το πιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μία ή k συστάδες. Λιγότερο διαδεδομένος είναι ο **Διαιρετικός τύπος**, όπου αρχίζουμε με μία συστάδα που περιέχει όλα τα σημεία και σε κάθε βήμα, διαχωρίζουμε μία συστάδα, μέχρι το σημείο όπου κάθε συστάδα να περιέχει μόνο ένα σημείο. Οι παραδοσιακοί αλγόριθμοι, χρησιμοποιούν έναν πίνακα ομοιότητας ή απόστασης, που παρακάτω θα αποκαλούμε **πίνακα Γειτνίασης**.

Η πιο δημοφιλής τεχνική συσταδοποίησης είναι η Συσσωρευτική Ιεραρχική Συσταδοποίηση (ΣΙΣ), όταν εφαρμόσουμε τη ΣΙΣ ακολουθούμε τα ακόλουθα βήματα:

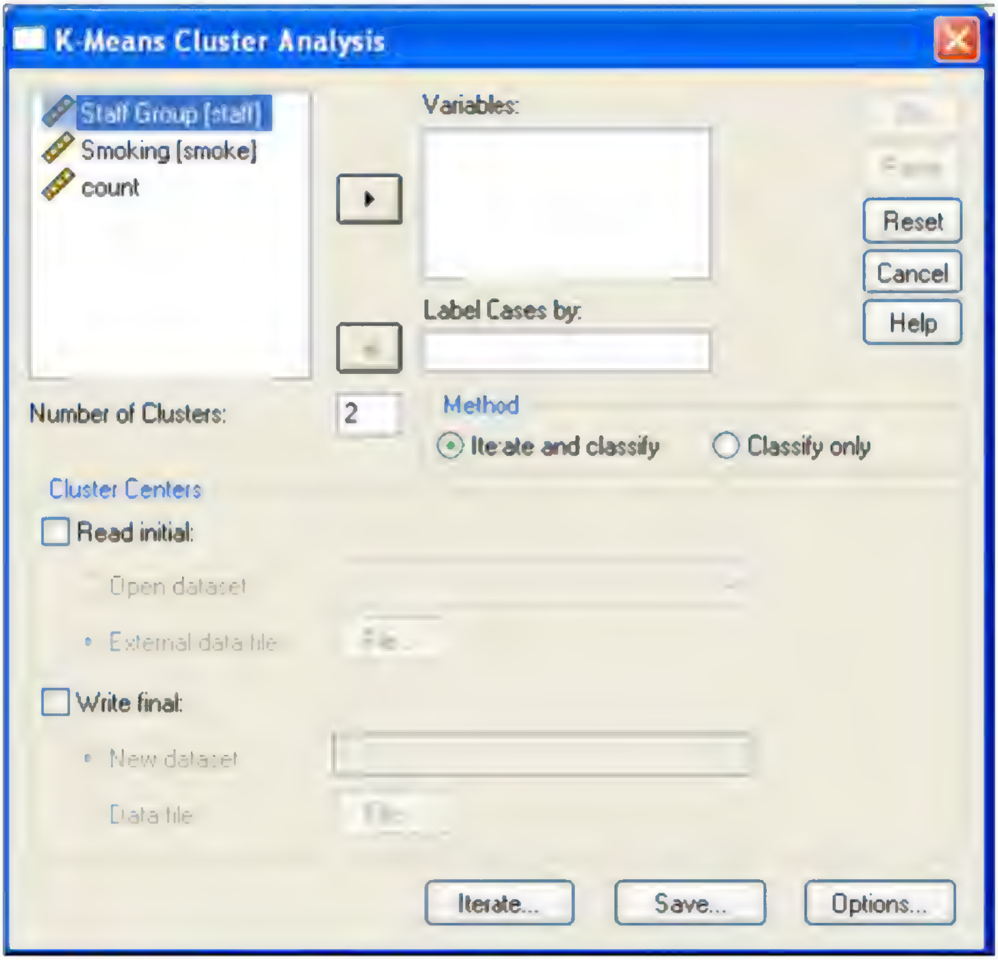
1. Υπολογίζουμε τον Πίνακα Γειτνίασης.
2. Θεωρούμε πως κάθε σημείο αποτελεί και μία συστάδα.
3. Επαναλαμβάνουμε την παραπάνω διαδικασία.
4. Συγχωνεύουμε τις δύο κοντινότερες συστάδες.
5. Ενημερώνουμε-ανανεώνουμε τον πίνακα Γειτνίασης.
6. Επαναλαμβάνουμε την παραπάνω διαδικασία μέχρι να μείνει μία μόνο συστάδα.

Στο σημείο αυτό θα πρέπει να επισημάνουμε, πως ο πίνακας Γειτνίασης είναι διαφορετικός κάθε φορά που χρησιμοποιούμε διαφορετική προσέγγιση για τον υπολογισμό της απόστασης.

1.4 Χρήση του SPSS για τις παραπάνω μεθόδους

Για να εκτελέσουμε την K-means ομαδοποίηση επιλέγουμε από το βασικό μενού του SPSS:

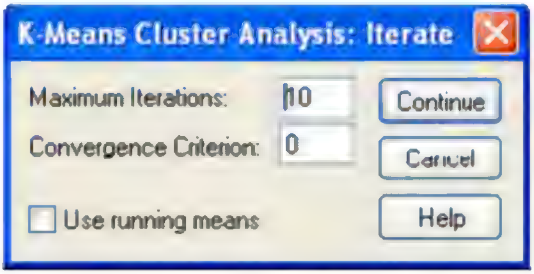
Analyze -> Classify -> K-means



Σχήμα 1.4.1: K-means ομαδοποίησης στο SPSS

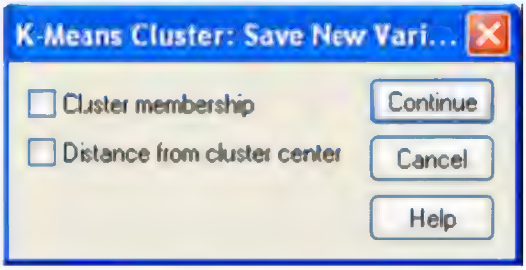
Στο παράθυρο διαλόγου "K-means Cluster Analysis" (Σχήμα 1.4.1) επιλέγουμε τις μεταβλητές που θα χρησιμοποιήσουμε για να υλοποιήσουμε τη συγκεκριμένη μέθοδο. Επιπλέον, στην ένδειξη "Number of cluster" εισάγουμε το πλήθος των ομάδων που επιθυμούμε. Από το συγκεκριμένο παράθυρο διαλόγου μπορούμε να ενεργοποιήσουμε τα ακόλουθα παράθυρα από τις επιλογές Iterate – Save – Options.

Από την επιλογή "Iterate" εμφανίζεται το παράθυρο διαλόγου "K-means Cluster Analysis Iterate" (Σχήμα 1.4.2), στο σημείο αυτό επιλέγουμε το πλήθος των επαναλήψεων που θα πραγματοποιήσει ο αλγόριθμος μέχρι να τερματίσει (Maximum Iterations) και τη μεγαλύτερη απόσταση ανάμεσα σε διαδοχικά κέντρα όλων των ομάδων (Convergence Criterion).



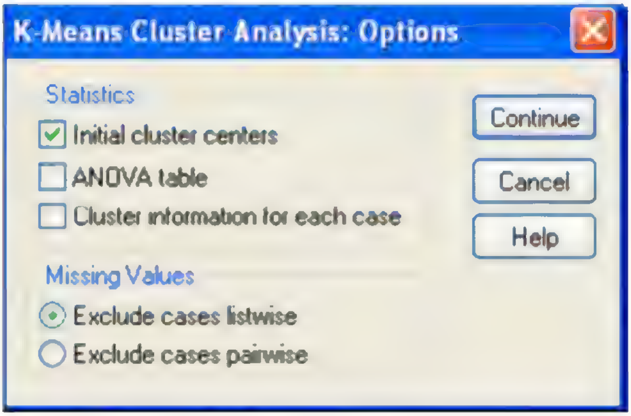
Σχήμα 1.4.2: Παράθυρο διαλόγου για τα κριτήρια τερματισμού

Από την επιλογή "Save", εμφανίζεται το παράθυρο διαλόγου "K-means Cluster Save New Variable" (Σχήμα 1.4.3). Με την επιλογή "Cluster Membership" δημιουργείται μια νέα στήλη, η οποία αντιστοιχεί σε κάθε παρατήρηση μία τιμή που χαρακτηρίζει την ομάδα που την κατατάξαμε, με αποτέλεσμα να μπορούμε να δούμε τα χαρακτηριστικά κάθε ομάδας.



Σχήμα 1.4.3: K-means Cluster Save New Variable

Από την τελευταία επιλογή "Options", εμφανίζεται το παράθυρο διαλόγου "K-means Cluster Analysis Options" (Σχήμα 1.4.4). Στο στάδιο αυτό επιλέγουμε ποια αποτελέσματα επιθυμούμε να εμφανιστούν. Έχουμε τη δυνατότητα, να εμφανιστούν τα αρχικά κέντρα, για να έχουμε πιο λεπτομερή εικόνα, για τις διαδοχικές εκτελέσεις του αλγόριθμου.



Σχήμα 1.4.4: K-means Cluster Analysis Options

Οι πίνακες αποτελεσμάτων, που προκύπτουν, είναι οι ακόλουθοι:

Initial Cluster Centers: Περιέχει τα αρχικά κέντρα των ομάδων.

ANOVA: Στο συγκεκριμένο πίνακα ελέγχουμε, για το αν διαφέρουν οι μέσες τιμές ανάμεσα στις ομάδες. Συνεπώς, οι μεταβλητές, που έχουν ικανότητα να διαχωρίζουν τις παρατηρήσεις εμφανίζονται στατιστικά σημαντικές. Σε αυτό το σημείο είναι σκόπιμο να επισημάνουμε, πως ο αλγόριθμος έχει κατασκευαστεί με τέτοιο τρόπο ώστε κάθε φορά να μεγιστοποιεί την ελγχοσυνάρτηση, συνεπώς η χρήση των αποτελεσμάτων είναι καθαρά περιγραφική.

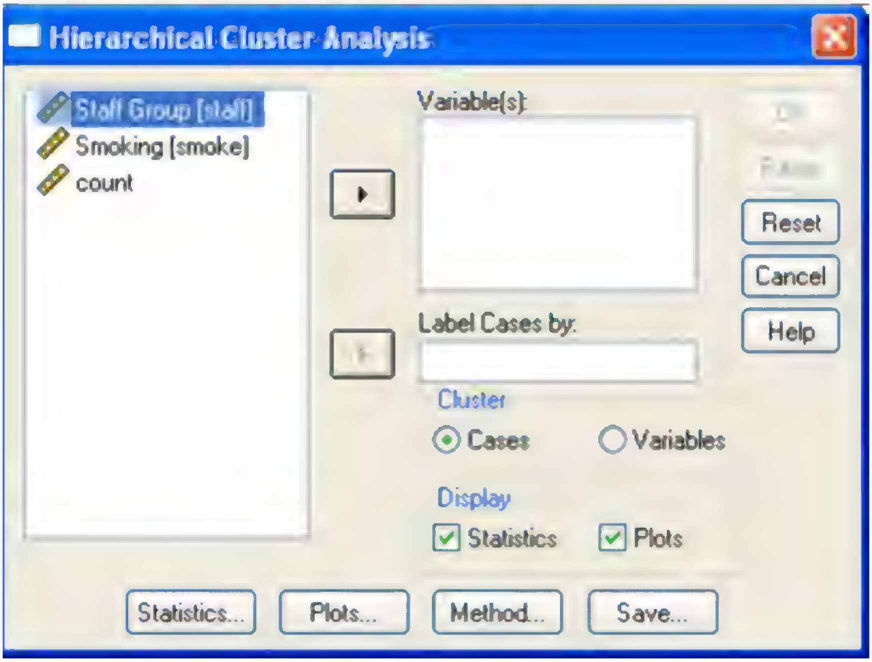
Iteration History: Περιέχει πληροφορίες για τις αλλαγές του αλγόριθμου σε κάθε επανάληψη.

Final Cluster Centers: Περιέχει τα τελικά κέντρα των ομάδων που βρέθηκαν αφού τερματίστηκε ο αλγόριθμος.

Number of Cases in each Cluster: Ο πίνακας παρουσιάζει πόσες παρατηρήσεις περιέχει κάθε ομάδα τελικά.

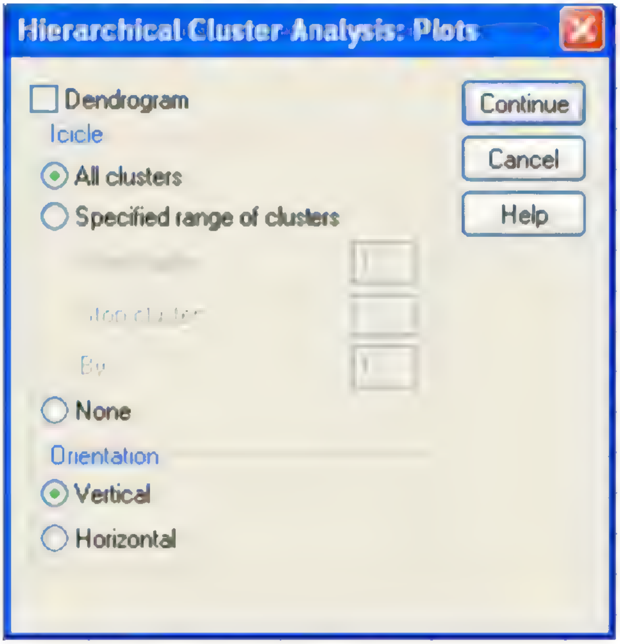
Για να εκτελέσουμε την ιεραρχική ομαδοποίηση στο SPSS διαλέγουμε:

Analyze -> Classify -> Hierarchical Clustering



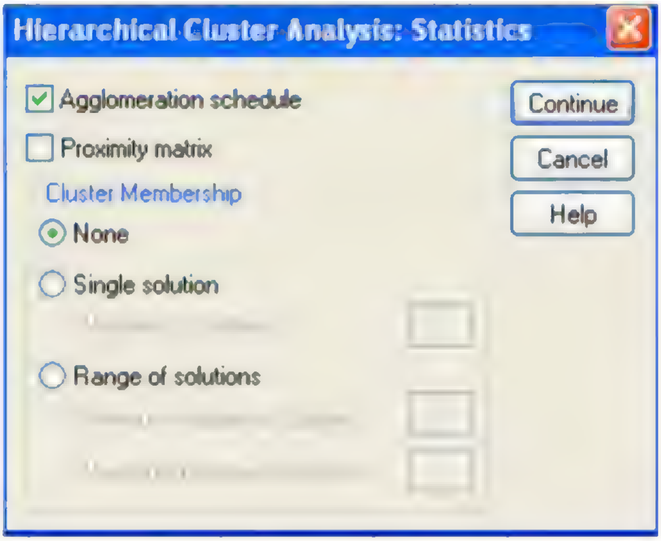
Σχήμα 1.4.5: Ιεραρχική ομαδοποίηση στο SPSS

Στο παράθυρο διαλόγου "Hierarchical Cluster Analysis" (Σχήμα 1.4.5) επιλέγουμε τις μεταβλητές που θα χρησιμοποιήσουμε στην ανάλυση. Από την επιλογή "Plots" εμφανίζεται το ακόλουθο παράθυρο.



Σχήμα 1.4.6: Hierarchical Cluster Analysis: Plots

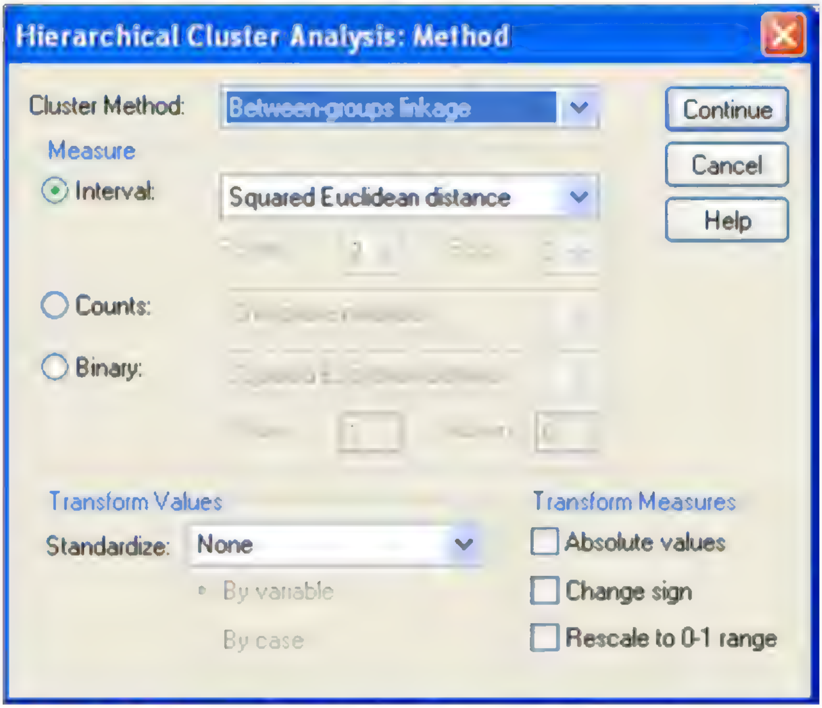
Επιλέγοντας την ένδειξη "Dendrogram", θα εμφανιστεί στα αποτελέσματά μας το δενδρόγραμμα, που θα μας απεικονίσει γραφικά τη σειρά με την οποία οι παρατηρήσεις ενώνονται για να δημιουργηθούν οι ομάδες. Το μειονέκτημα που παρουσιάζει η συγκεκριμένη απεικόνιση είναι πως τα γραφήματα δεν είναι ευανάγνωστα στην περίπτωση που ο αριθμός των παρατηρήσεων είναι αρκετά μεγάλος. Σε αυτή την περίπτωση έχουμε τη δυνατότητα να επιλέξουμε το εύρος του αριθμού των ομάδων για τις οποίες θα εμφανιστεί το γράφημα (Specified range of cluster) και να ορίσουμε αν το γράφημα θα εμφανιστεί οριζόντια (Horizontal) ή κάθετα (Vertical).



Σχήμα 1.4.7: Hierarchical Cluster Analysis: Statistics

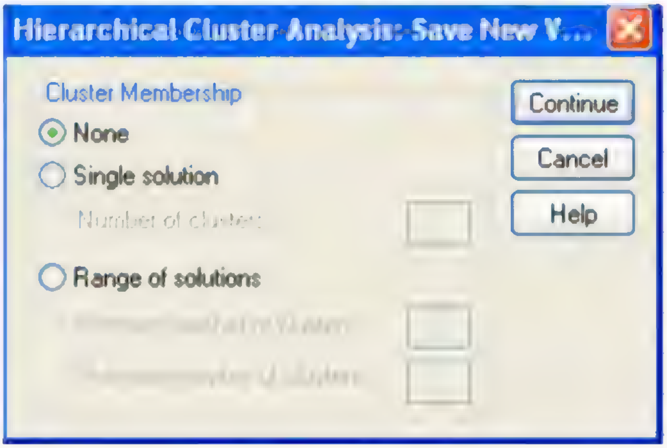
Από το παράθυρο διαλόγου του Σχήματος 1.4.5, επιλέγοντας την ένδειξη "Statistics", εμφανίζεται το παραπάνω παράθυρο (Σχήμα 1.4.7). Επιλέγοντας "Agglomeratiion Schedule", εμφανίζονται οι ποσότητες, που θα μας φανούν χρήσιμες για να καταλήξουμε στον αριθμό των ομάδων που θα καταλήξουμε. Με την "Proximity Matrix", θα έχουμε στα αποτελέσματά μας τον πίνακα αποστάσεων όλων των παρατηρήσεων.

Επανερχόμαστε στο παράθυρο διαλόγου του Σχήματος 1.4.5 και επιλέγουμε την ένδειξη "Method". Στο συγκεκριμένο παράθυρο διαλόγου (Σχήμα 1.4.8) καθορίζουμε τη μέθοδο που θα χρησιμοποιήσουμε για να υπολογίσουμε την απόσταση ανάμεσα σε δύο ομάδες.



Σχήμα 1.4.8: Hierarchical Cluster Analysis: Methods

Τέλος, από το παράθυρο διαλόγου του Σχήματος 1.4.5, επιλέγοντας την ένδειξη "Save", εμφανίζεται το παρακάτω παράθυρο (Σχήμα 1.4.9), όπου μπορούμε να δημιουργήσουμε μεταβλητές που να μας δείχνουν, για τη συγκεκριμένη λύση με το συγκεκριμένο αριθμό ομάδων. Επιπλέον, αν επιθυμούμε μπορούμε να δημιουργήσουμε μεταβλητές για πολλές δυνατές λύσεις ανάλογα με τον αριθμό των ομάδων.

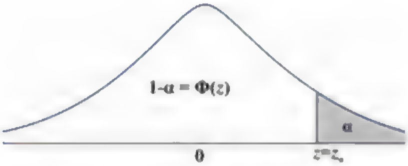


Σχήμα 1.4.9: Hierarchical Cluster Analysis: Save New Variables

ΠΑΡΑΡΤΗΜΑ III

Πίνακας 1

Τιμές των πιθανοτήτων $\Phi(z)=P(Z\leq z)$ της τυποποιημένης κανονικής κατανομής $N(0,1)$ για $z\geq 0$. Για $z<0$ ισχύει $\Phi(z)=1-\Phi(-z)$.

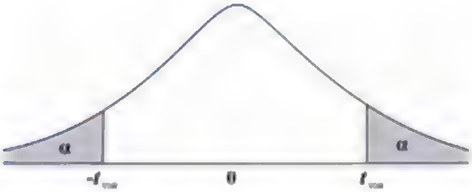


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84850	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92786	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900

α	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
z_α	3.29	3.09	2.576	2.326	1.960	1.645	1.282

Πίνακας 2

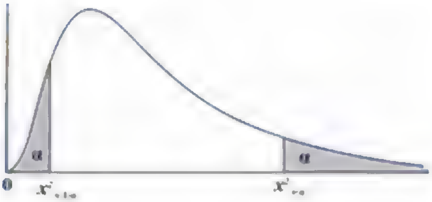
Των τιμών $t_{v,a}$ της t_v-κατανομής ώστε $P(t^*_v > t_{v,\alpha}) = P(t^*_v \geq t_{v,\alpha}) = \alpha$.



v	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Πίνακας 3

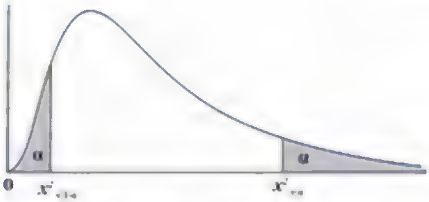
Των τιμών $\chi^2_{v,1-a}$ της χ^2 κατανομής για τις οποίες $P(X < \chi^2_{v,1-a}) = P(X \leq \chi^2_{v,1-a}) = a$.



2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Πίνακας 3 (συνέχεια)

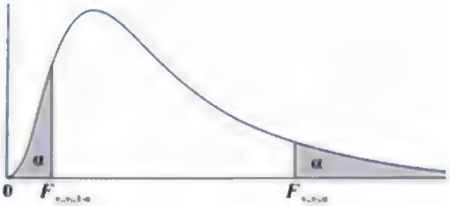
Των τιμών $\chi^2_{v,a}$ της χ^2 κατανομής για τις οποίες $P(X > \chi^2_{v,a}) = P(X \geq \chi^2_{v,a}) = a$.



v	a = 0.10	a = 0.05	a = 0.025	a = 0.01	a = 0.005
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21,9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169

Πίνακας 4

Τιμές της $F_{v1,v2,\alpha}$ της F κατανομής για τις οποίες $P(X>F_{v1,v2,\alpha})=P(X\geq F_{v1,v2,\alpha})=\alpha(\alpha=0,01)$.
Για τα α-κάτω ποσοστιαία σημεία $F_{v1,v2,1-\alpha}$ ισχύει η σχέση $F_{v1,v2,1-\alpha}=1/F_{v2,v1,\alpha}$.

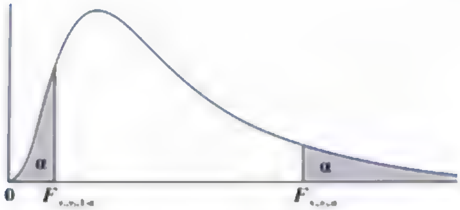


ν_1 / ν_2	1	2	3	4	5	6	7	8	9
1	4.052	4.9995	5.403	5.625	5.764	5.859	5.928	5.982	6.022
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Πίνακας 4 (συνέχεια)

Τιμές της $F_{v1,v2-a}$ της F κατανομής για τις οποίες $P(X>F_{v1,v2-a})=P(X\geq F_{v1,v2-a})=a(a=0,01)$.

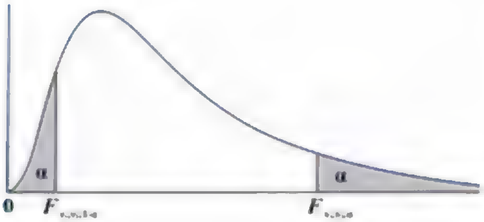
Για τα α-κάτω ποσοστιαία σημεία $F_{v1,v2,1-a}$ ισχύει η σχέση $F_{v1,v2,1-a} = 1/ F_{v2,v1,a}$.



ν_1 / ν_2	10	12	15	20	24	30	40	60	120	∞
1	6.056	6.106	6.157	6.209	6.235	6.261	6.287	6.313	6.339	6.366
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Πίνακας 4 (συνέχεια)

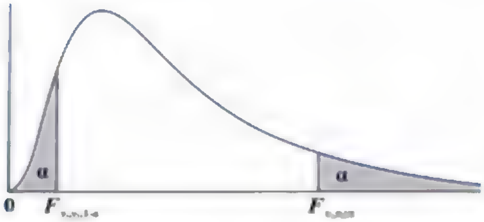
Τιμές $F_{v1,v2,a}$ της F κατανομής για τις οποίες $P(X>F_{v1,v2,a})=P(X\geq F_{v1,v2,a})=a(a=0,05)$.
Για τα a-κάτω ποσοστιαία σημεία $F_{v1,v2,1-a}$ ισχύει η σχέση $F_{v1,v2,1-a} = 1/F_{v2,v1,a}$.



v_1 / v_2	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Πίνακας 4 (συνέχεια)

Τιμές $F_{v1,v2,a}$ της F κατανομής για τις οποίες $P(X>F_{v1,v2,a})=P(X\geq F_{v1,v2,a})=a(a=0,05)$.
Για τα α-κάτω ποσοστιαία σημεία $F_{v1,v2,1-a}$ ισχύει η σχέση $F_{v1,v2,1-a} = 1/ F_{v2,v1,a}$.



ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Αθανασόπουλος Δ., «Επαγωγική Στατιστική», Εκδόσεις Σταμούλης, 1990.
- Αθανασόπουλος Δ., «Θεωρία Πιθανοτήτων 1 & 2», Εκδόσεις Σταμούλης, 1991.
- Αθανασόπουλος Δ. και Μπένος Β., «Εφαρμογές Στατιστικής», Τόμοι Α΄ & Β′ Εκδόσεις Σταμούλης, 1990.
- Δρακάτος Κ. Γ., «Στατιστική», Εκδόσεις Παπαζήση, 1984.
- Δρακάτος Κ. Γ., «Ασκήσεις Στατιστικής», Εκδόσεις. Παπαζήση, Αθήνα, 1984.
- Ζαχαροπούλου Χ., «Στατιστική, Μέθοδοι – Εφαρμογές» Τόμοι Α′ & Β′, Εκδόσεις ΣΟΦΙΑ Α.Ε., 2005 & 2008, αντίστοιχα.
- Ζαχαροπούλου Χ., «Ασκήσεις Στατιστικής», Εκδόσεις ΣΟΦΙΑ Α.Ε., 2002.
- Θαλασσινός Ε., Σταματόπουλος. Θ., και Χαρίσης Χ., «Επιχειρησιακή Στατιστι-κή», Εκδόσεις Σταμούλης, 1996.
- Ιωαννίδης Δ. Α., «Στατιστικές Μέθοδοι», Τόμος Ι, Εκδόσεις Ζήτη, Θεσσαλονίκη, 2001.
- Καλαματιανού Α. Γ., «Κοινωνική Στατιστική: Μέθοδοι Μονοδιάστατης Ανάλυ-σης», Εκδόσεις Οικονομικό, Αθήνα, 1992.
- Κιντής Α., «Σύγχρονη Στατιστική Ανάλυση», Εκδόσεις Gutenberg, Αθήνα, 1995.
- Κιντής Α., «Στατιστικές και Οικονομετρικές Μέθοδοι», Εκδόσεις Gutenberg, Α-θήνα, 1994.
- Κάτος Α. Β., «Στατιστική», Εκδόσεις Παρατηρητής, Θεσσαλονίκη, 1986.
- Λαμπράκης Δ., «Στατιστική», Αυτοέκδοση, 1980.
- Λουκάς Σ. Β., «Στατιστική», Εκδόσεις Κριτική, 2003.
- Μπένος Β., «Εφαρμογές Επαγωγικής Στατιστικής με Στοιχεία Θεωρίας», Εκδόσεις Σταμούλης, 1997.
- Μπένος Β., «Μέθοδοι και Τεχνικές Δειγματοληψίας», Εκδόσεις Σταμούλης, 1991.
- Μπένος Β., Κούτρας Μ. και Αντζουλάκος Δ., «Ασκήσεις Πιθανοτήτων», Μέρος Ι, Εκδόσεις Σταμούλης, 2004.
- Παπαδημητρίου Γ., «Περιγραφική Στατιστική», Εκδόσεις Τυπωθήτω,2005.
- Πανάρετος Ι., Ξεκαλάκη Ε., «Εισαγωγή στη Στατιστική Σκέψη», τόμος Ι (Περι-γραφική Στατιστική), 1997.

- Πέκος Γ. Δ., «Ασκήσεις Στατιστικής», Εκδόσεις ΖΥΓΟΣ, Θεσσαλονίκη, 2006.
- Τερζάκης Δ., «Στατιστική Επιχειρήσεων, με Εφαρμογές στον Τομέα του Τουρι-σμού», Εκδόσεις Interbooks, 1999.
- Χαλικιάς Ι., «Στατιστική: Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις», Εκδόσεις Rossili, Αθήνα, 2001.
- Χατζηνικολάου Δ., «Στατιστική για Οικονομολόγους», Ιωάννινα, 2002
- Cryer J. and Cobb G., «An Electronic Companion to Business Statistics», Edi-tion Oxford University Press, 1997.
- Groebner D., P.Shannon P.Fry, and K.Smith, «Business Statistics: A Decision-Making Approach», Edition Pearson, 2005.
- Jarret J., «Μέθοδοι Προβλέψεων για Οικονομικές - επιχειρηματικές Αποφάσεις», Εκδόσεις Gutenberg, Αθήνα, 1993.
- Johnson R. A., «Business Statistics: Decision Making with Data», Edition Wiley, 1997.
- Levine D., Stephan D., Krehbiel T. and Berenson M., «Statistics for Managers, using MS-Excel», 5th Edition Pearson, 2008.
- McClave J., Benson P. and Sincich T., «Statistics for Business and Economics», 9th Edition Prentice-Hall, 2005.
- Mendenhall W., Beaver R. J. and Beaver B. M., «Introduction to Probability & Statis-tics» 10th Edition Duxbury Press, 1999.
- Siegel A., «Practical Business Statistics», 3/e Edition IRWIN, 1996.
- Sincich T. «Business Statistics by Examples» 5th Edition Prentice-Hall, 1996.
- Sincich T. «Business Statistics by Example», Edition Maxwell MacMillan Co., Toronto, 1992.

ΗΛΕΚΤΡΟΝΙΚΕΣ ΔΙΕΥΘΥΝΣΕΙΣ

ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ «STATISTICA»
http://www.statsoft.com/

ΛΕΞΙΚΟ ΣΤΑΤΙΣΤΙΚΩΝ ΟΡΩΝ
http://www.statsoft.com/textbook/glosfra.html
http://www.statsoft.com/textbook/stathome.html
http://www.stat.ufl.edu

ΣΤΑΤΙΣΤΙΚΗ ΒΙΒΛΙΟΘΗΚΗ
http://davidmlane.com/hyperstat/
http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html
http://home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm

ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ «MINITAB»
http://www.minitab.com/
http://www.pitt.edu/~nancyp/stat-0200/minitab15.html
http://www.saintmarys.edu/~cpeltier/MTBhbk/mtbindex.html http://www.ma.utexas.edu/restricted-resources/utma-doc/minitab/minitab.html

ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ «SPSS»
http://www.spss.com/
http://www.duxbury.com/statistics_d/templates/student_resources/0534391869/bkSite/spss/index.html
http://comp.uark.edu/~whlevine/psyc2013/
www.hop.man.ac.uk/Academic/researchdevelopment/statsdocs/SPSSnote.doc
http://homepages.ed.ac.uk/ercn82/drafts/
http://www.spsstools.net/

ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ «SPLUS»
http://www.splus.com/
http://www.math.unm.edu/splus/Splus.html
http://csg.sph.umich.edu/docs/R/S-tutorial.pdf
http://web.nps.navy.mil/~buttrey/SIndex.html
http://www.statslab.cam.ac.uk/~pat/
http://www.r-project.org/
http://www.idrc.ca/en/ev-56449-201-1-DO_TOPIC.html

ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ
http://www.probability.net/
http://www.richland.edu/james/lecture/m170/
http://mathforum.org/dr.math/faq/faq.prob.intro.html
http://www-math.bgsu.edu/~albert/m115/probability/outline.html
http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html
http://www.ipp.mpg.de/de/for/bereiche/stellarator/Comp_sci/CompScience/csep/csep1.phy.ornl.gov/mc/node4.html

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

http://www.gonzaga.edu/doctoral/id722/id722-1/M1PM.html
http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage22.htm
http://mathworld.wolfram.com/topics/DescriptiveStatistics.html
http://cas-courses.buffalo.edu/classes/psy/segal/1slideset/index.html
http://www.nt2.sphtm.tulane.edu/courses/biostat603/chap1.htm

ΕΚΤΙΜΗΤΙΚΗ – ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage26.htm
http://forrest.psych.unc.edu/research/vista-frames/help/class-notes.html
http://cas-courses.buffalo.edu/courses/psy/segal/sampdist/sampdist2.htm
http://cas-courses.buffalo.edu/courses/psy/segal/sampdist/sampdist2.htm

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΙ ΣΧΕΔΙΑΣΜΟΣ ΠΕΙΡΑΜΑΤΩΝ

http://www.ruf.rice.edu/%7Emickey/psyc339/notes/ANOVA.html
http://www.ruf.rice.edu/%7Emickey/psyc339/notes/ANOVA2.html
http://www.gonzaga.edu/doctoral/Id722/Id722-5/M5PM.html
http://trochim.human.cornell.edu/kb/design.htm
http://trochim.human.cornell.edu/kb/statfact.htm
http://trochim.human.cornell.edu/kb/statblck.htm
http://trochim.human.cornell.edu/kb/statcov.htm
http://cas-courses.buffalo.edu/classes/psy/segal/ANOVA1/Anova.html
http://cas-courses.buffalo.edu/courses/psy/segal/anova2/ANOVA2.html

ΕΙΣΑΓΩΓΗ ΣΤΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

http://www.biometrics.mtu.edu/FW5411.html
http://elsa.berkeley.edu/sst/regression.html
http://www.amstat.org/publications/jse/v2n2/laviolette.html
http://www-stat.stanford.edu/~jtaylo/courses/stats203/index.html
http://www.stat.tamu.edu/stat30x/notes/trydouble2.html

ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

http://www.statsoft.com/textbook/stathome.html
http://statistics.byu.edu/faculty/wfc/stat435/
http://sst-web.tees.ac.uk/external/U0000504/Notes/DataAnalysis/DistFree/NonParaExercise.html
http://forrest.psych.unc.edu/research/vista-frames/help/lecturenotes/lecture09/lec9part4.html
http://www.math.wustl.edu/~sawyer/math408s05.html
http://legacy.ncsu.edu/classes/psy242_243001/wilcoxin/WILCOXIN.html

ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

http://www.agu.org/revgeophys/lall01/node4.html
www.stat.auckland.ac.nz/~ihaka/726/notes.pdf
http://www.qmw.ac.uk/~ugte133/courses/tseries/rootseri.html

ΤΕΧΝΙΚΕΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

http://drott.cis.drexel.edu/sample/content.html
http://www.willyancey.com/sampling-theory.html#stratification